European research infrastructures (including e-Infrastructures)

Topic: Implementing the European Open Science Cloud

European research infrastructures (including e-Infrastructures)

FAIR Mathematical Data for the European Open Science Cloud

Acronym: *FAIRMat*

Date of Preparation: Tuesday 29th January, 2019: 16:52

Coordinator: Michael Kohlhase

e-mail: michael.kohlhase@fau.de

tel/fax: +49 9131-85-64052/55

Keywords: mathematics, computer algebra, simulation, datasets

#	Participant organisation name	Short name	Country
1	Friedrich-Alexander Universität Erlangen/Nürnberg (coordinator)	FAU	DE
2	Université Paris-Sud	UPSud	FR
3	Chalmers University of Technology	CHA	SE
4	Univerza v Ljubljani	UL	SI
5	CAE Tech Limited	CAE	UK
6	FIZ Karlsruhe – Leibniz Institute for Information Infrastructure	FIZ	DE
7	European Mathematical Society	EMS	FI

Abstract

The scientific community increasingly considers datasets as their own kind of resource that should be shared and published individually, either in conjunction with a traditional paper or as a standalone digital artefact. Recently this trend has formed a positive feedback loop with the rise of deep learning methods, which require large datasets as input. Multiple national and international Open Science initiatives have been started to ensure the open availability, easy sharing, and reliable reproducibility of datasets in particular and data-driven research in general. The FAIR principles in particular have been developed as a goalpost for the openness of datasets, including in particular the sharing across disciplines and between research and industry.

As a rule, mathematicians strongly support the Open Science movement and happily make their datasets public for both practical and ethical reasons. This is accompanied by a vibrant and growing community of Open Source software for computational mathematics. However, the systematic FAIR sharing of mathematical data is very difficult due to the inherent complexity of the data. Therefore, today most mathematical data collections are shared in an ad hoc manner that is limited in scope and suffers from a lack of interlinking of digital artefacts across platforms. Thus, FAIR mathematics, while widely welcomed, is effectively non-existent today. A similar argument applies to related sciences to the extent that they make heavy use of mathematical data, e.g., the mathematical modeling of cyber-physical systems.

Generally, reusing shared data requires that the reuser be able to understand the semantics of the data. This is particularly difficult for system interoperability where the semantics must not only be evident but must itself be accessible for automated processing, and it is particularly critical where data is used in safety-critical systems. While this problem exists for all data, it is particularly challenging for mathematical and similar data in related disciplines where the semantics is very difficult to specify. Therefore, today there are virtually no mathematical datasets whose semantics is itself accessible.

This is evidenced by the wide gap in the service offerings of the EOSC when it comes to semantics-aware services in general and services for mathematical data in particular.

FAIRMat (pronounced “Fermat”) will deliver a framework and prototype service for the FAIR and semantics-aware sharing of mathematical data. It will meet the needs of mathematics research and education and will also bring added value to other disciplines and industry that work with data that exhibits complex structure or semantics. It will support all phases of the research life-cycle including the generation, publication, updating, extending, curation, search, reuse, and archival of mathematical data.

FAIRMat will be based on a sustainable software ecosystem including Open Source databases and services as well as nationally funded infrastructures. This includes TRL 6 and above technologies like the Modelica modeling language, the LMFDB database, the SageMath computer algebra suite, or the zbMATH publication information system.

Besides supporting the direct sharing of newly-produced data, *FAIRMat* will make existing datasets available in a uniform way. In particular, we will demonstrate the scalability of the *FAIRMat* infrastructure by integrating it with several large and important databases and services from diverse communities including databases of pure mathematical objects (graphs, integer sequences, elliptic curves, . . .), formalized theorems and proofs, mathematical models used in engineering, linked data from Wikidata, and publication metadata. Finally, to maximize its long-term impact, *FAIRMat* is designed to culminate in the ISO standardization of its data representation format and the smooth integration of its services into the EOSC Hub at the end of the project lifetime.

FAIRMat will be carried out by a consortium of 7 sites with huge experience in designing and maintaining mathematical datasets and services. The majority of partners are long time Open Science promoters, with a strong experience in large open (software) project management. In particular, all produced documents (including this proposal itself), software, and data will be licensed under open licenses and freely available.

Contents

1 Excellence **3**
1.1 Objectives . . . 6
1.2 Relation to the Work Programme . . . 8
1.3 Concept and Methodology . . . 9
1.4 Ambition . . . 17

2 Impact **26**
2.1 Expected Impacts . . . 26
2.2 Measures to Maximize Impact . . . 29

3 Implementation **33**
3.1 Work Plan — Work packages, deliverables . . . 33
3.2 Management Structure, Milestones, and Procedures . . . 52
3.3 Consortium as a Whole . . . 56
3.4 Resources to be Committed . . . 58

4 Members of the Consortium **65**
4.1 Participants . . . 65
4.2 Third Parties Involved in the Project (including use of third party resources) . . . 82

5 Ethics and Security **83**
5.1 Ethics . . . 83
5.2 Security . . . 83

1 Excellence

Mathematics as a Motor for Innovation: Innovations based on mathematical knowledge and algorithms yield many improvements in economy, ecology, health care, security, and in society overall. Our global positioning system (GPS) needs the mathematics of relativistic physics, our mobile phones use frequencies allocated through combinatorial optimization, the combinatorics of our genome yields clues to curing rare diseases, the privacy of our communications depends on cryptographic protocols steeped in number theory, and our national security is relying on the mathematical analysis of increasingly complex networks. Fundamental mathematical research and its direct application in practical situations enable many engineering, science, and business innovations that enrich society and mankind. Such applications more and more drive modern mathematical research, which depends critically and increasingly on collaborative tools, computational environments, and online databases. Many of these digital tools have revolutionized the way mathematical research is conducted and how it is turned into applications. For example, engineers now use mathematical tools to build and simulate physical models based on systems of differential equations and using millions of variables, combining building blocks and algorithms taken from libraries from all over the internet.

Problem: Oligopolization of Mathematics There is very high commercial interest in the development of mathematical representations as proprietary services and datasets, which leads to the danger of monopolizing their availability. Indeed we are seeing that the large engineering and internet companies are strategically buying (all) the relevant, innovative startups and hiring top researchers, essentially privatizing and oligopolizing public data, knowledge, and technological know-how. Even in the field of mathematics — which could be assumed to be "pure" and thus immune — this is the case for, e.g., machine learning algorithms or the data curation of Wolfram Inc., which has started integrating mathematical data into the Wolfram Language almost a decade ago.

The Cure: Open Data/Software: We are strongly convinced that mathematical data and algorithms should be openly available for the research community and industry according to the FAIR principles [FAIR18] and that there should be open access to all resources. The members of this consortium have demonstrated this commitment and its benefits with the open access services and datasets they have developed or contributed to in the past, such as Modelica [MOb], SLICOT [Ben+99], SageMath [Sage], EuDML [EUD], swMATH [SWM], or LMFDB [LM].

Project Aim: We will provide mathematicians and scientists with *i*) a standardized unified framework for representing mathematical data, *ii*) a scalable set of user-oriented FAIR services for them, and *iii*) integrate with the resulting software infrastructure a collection of existing datasets that are widely used in the mathematical community.

Prerequisite: Deep FAIRness: To achieve this, we will need to build services that understand the semantics [FAIR18, Rec. 7] of the mathematical data they operate on — only if the mathematical meaning of the data is accessible in all its depth can computer applications provide mathematically sound, interoperable services. We call this *deep* FAIRness.

Due to the mathematical standard of rigor and the inherent complexity of mathematical data, deep FAIRness is both more difficult and more important for mathematics than for other scientific disciplines. That also means that mathematics is an ideal test case for developing the semantic aspects of the FAIR principles in general.

A lot of FAIR-motivated knowledge sharing has already been done. A few examples from different walks of mathematics are:

i. The Modelica language uses symbolic representations of differential equations and control algorithms to model cyber-physical systems and bases simulation services on that. Hundreds of reusable libraries are available on GitHub alone.

ii. Highly standardised subroutine libraries like LAPACK [LPK], SLICOT, or MUMPS [MPS] form the backbone of almost all engineering software packages.

iii. Mathematical information services like zbMATH [ZBM], EuDML, and swMATH extend bibliographic metadata of mathematical publications with math subject classifications (essentially taxonomic semantic information) and use automatic extraction to give users enhanced, semantic search capabilities.

iv. Libraries of formalized mathematics directly specify the meaning of mathematical definitions, theorems, and proofs in a machine-verifiable way. Tens of thousands of such formal proofs are available in open libraries.

v. Mathematical databases like the L-functions and Modular Forms DataBase (LMFDB [LM]), the GAP Small Groups Library [EBO], or the Open Encyclopedia of Integer Sequences (OEIS [OEIS]) store millions of mathematical objects together with their semantic properties, both human-curated or machine-generated.

These are used industrially (*i.*-*iv.*) and academically (*i.*-*v.*), and inner-mathematically (*ii.*, *iv.*, *v.*) and transdisciplinarily (e.g. *i.*-*ii.* in engineering and *iii.* in program verification). But the various representations are non-interoperable, and the datasets therefore are not reusable across systems and communities. This leads to large gaps in the FAIRness of mathematical data and results in missed opportunities for innovative services that could revolutionize mathematical research and applications.

Open Source/Data Ethos: The mathematical community predominantly shares the ethos of open access to publications, software (including source code), and datasets. In fact, all of the examples above are either fully open, partly open, or are currently in the process of opening up the data/software further. For mathematical software, the Open Source ethos has been established already for more than 50 years in subroutine libraries such as LAPACK, which are produced according to a widely accepted documentation and implementation standard and are at the core of almost all successful commercial and non-commercial software packages including MATLAB [MLB] and SageMath [Sage].

Throughout this project we will reuse and extend open source code, and *FAIRMat* will benefit from future open source contributions during and beyond the lifetime of the project. Moreover, the *FAIRMat* project will follow the example of the H2020 OpenDreamKit project and conduct all of its development openly in public repositories. In fact like the OpenDreamKit project, which *FAIRMat* follows up on, the *FAIRMat* proposal was developed publicly (on `https://gl.kwarc.info/mathhub/data-proposal/`). Thanks to this by-users-for-users model, *FAIRMat* will be steered by the actual needs of the community.

The *FAIRMat* team is a Europe-wide collaboration that brings together a leading body of mathematicians and transdisciplinary computational researchers, with an extensive track record of delivering innovative open source software solutions.

Impact: Standardizing a data framework, unifying services, and hosting all on a public, high-profile infrastructure like the EOSC will enable huge progress in effective research, research communication, and reproducibility in computational mathematics and related sciences. By focusing on public, open standards and service interoperability *FAIRMat* will simultaneously maximize sustainability and impact. Even though the primary target users are researchers in mathematics, the set of beneficiaries extends to researchers, teachers, and industry practitioners in, e.g., scientific computing, physics, chemistry, biology, engineering, medicine, earth sciences and geography, as well as social sciences and finance. *FAIRMat* will foster the development of models that are mutually beneficial to academia and highly innovative SMEs and enable tool chains that bridge the gap between fundamental mathematical research and domain-specific computational technology, thus supporting the faster application, exploitation, and commercialization of basic research.

Sustainability The result of the *FAIRMat* project will be a software infrastructure consisting of *i*) a uniform data representation standard that is ready for semantics-aware FAIR data sharing, *ii*) innovative user-oriented services that validate, serve, compute with, and visualize this data and leverage it in existing widely used applications, and *iii*) the uniform integration of multiple community-driven mathematical datasets that jointly comprise multi-terabyte databases. To make these results sustainable beyond the project duration, we will submit the above standard for ISO certification, develop all services in a way that allows for easy deployment on the EOSC Hub, and exploit the datasets made available uniformly through *FAIRMat* in extensive community outreach efforts to publicize the EOSC Hub and FAIR data sharing.

Moreover, the complete *FAIRMat* software infrastructure will run on a single, well-equipped, modern commodity-grade server that can be maintained sustainably by a part-time (1/4 FTE) experienced system administrator. In particular, the *FAIRMat* project is not applying for or incurring the future need of dedicated hardware infrastructure or large maintenance teams.

Funding Need: One might think that many of the solutions described above will eventually be organized by the mathematical community anyway. But a coordinated effort is needed to create a single, coherent data representation standard and the corresponding FAIR service framework. Without such a concerted effort — which requires the funding and institutional support as provided by the EOSC — we will likely see the continued development of a multitude of non-interoperable system-specific “standards” and competing commercial offerings, which are already becoming more and more entrenched. As mathematics is a small — albeit foundational — discipline, the *FAIRMat* proposal stays well below the recommended funding level provided in the INFRAEOSC-02-2019 call in order to make the proposed project cost-effective.

The time for the *FAIRMat* project is ideal as it is a follow-up to the successful OpenDreamKit project (2015–2019), which has built A Virtual Research Environment Toolkit for Mathematics. OpenDreamKit has identified many

of the problems and designed many of the solutions described in this proposal. Several of the *FAIRMat* proposers are OpenDreamKit alumni, and we expect being able to hire qualified OpenDreamKit personnel.

1.1 Objectives

1.1.1 Objective 1: FAIR Mathematics

Overview Traditionally, mathematics has not paid particular attention to the creation and sharing of data — the careful computation and publication of logarithm tables is a typical example of the extent and method. This has changed with the advent of computer-supported mathematics, and the practice of modern mathematics is increasingly data-driven. Today it is routine to use mathematical datasets in the Gigabyte range, including both human-curated and machine-produced data. Examples include the L-Functions and Modular Forms Data Base (LMFDB; ∼1TB data in number theory) [Cre16; LM] and the GAP Small Groups Library [EBO] with ∼450 million finite groups. In some — currently limited, but increasing — areas, mathematics has even become similar to experimental sciences in that mathematical reality is "measured" at large scale by running computations.

There is wide agreement in mathematics that these datasets should be a common resource and be open and freely available. Moreover, the software used to produce them is usually open source and free as well. Such an ecosystem is embraced by the mathematics community as a general vision for their future research infrastructure [Cou14], adopted by the International Mathematical Union as the Global Digital Mathematics Library initiative [GDML].

However, the datasets are produced, published, and maintained with virtually no systematic attention to the FAIR principles [FAIR18; Wil+16] for making data findable, accessible, interoperable, and reusable. In fact, often the sharing of data is an afterthought — see [Berb] for an overview of mathematical datasets and their "FAIR-readiness".

Moreover, the inherent complexity of mathematical data makes it very difficult to share in practice: even freely accessible datasets are often very hard or impossible to reuse, let alone make machine-interoperable because there is no systematic way of specifying the relation between the raw data and its mathematical meaning. Therefore, **unfortunately FAIR mathematics essentially does not exist today** — to mince no words: *mathematical data is currently largely unFAIR*.

The main objective of the *FAIRMat* project is to systematize the way how mathematical data is represented and shared in a way that enables FAIR mathematics. We will achieve this by standardizing a framework that allows both the representation of the various forms of mathematical datasets and the formal specification of their semantics. This will enable us to integrate a variety of mature services and datasets FAIRly.

Secondary Benefits A standard for FAIR data representations in mathematics will lead to several incidental benefits:

- It increases the productivity of mathematicians by allowing them to focus on the mathematical datasets themselves while leaving issues of encoding, management, and search to dedicated systems.
- It improves the reliability of published results as the research community can more easily scrutinize the underlying data.
- It enables collaborations via shared datasets that are currently prohibitively expensive due to the difficulty of understanding other researchers' data. More generally, it makes mathematical datasets available to researchers from other disciplines and industry practitioners, who are currently excluded due to the difficulty of understanding the datasets.
- It rewards mathematicians for sharing datasets (which is currently often not the case), e.g., by making datasets citable and their reuse known.
- It makes research more sustainable by guaranteeing that datasets can be archived and their meaning understood in perpetuity (which is essential especially in mathematics).

1.1.2 Objective 2: Semantics-Aware Open Science Cloud

We aim to leverage the semantics-aware framework for mathematical data developed in Objective 1 in two ways.

1. We build an EOSC-level software infrastructure for the FAIR sharing of mathematical datasets. This includes a suite of universal, scalable, and freely available services for finding, accessing, interacting with, and reusing in a semantics-aware way. Such services are currently non-existent, limited, or expensive.

2. We make a representative set of major existing datasets available through our infrastructure. These — and the servicers operating on them — cover the whole spectrum from bibliographic information to pure mathematics to industrial uses mathematical models of cyber-physical systems.

Our infrastructure will be realized as a self-contained prototype server with both a machine-oriented API and a human-oriented web interface. All services will be open to the public, free of charge and the contents licensed with suitable Open Source/Open Data licenses (see **T2.3**). We will also take all necessary steps to make the eventual integration into the EOSC hub a straightforward process that does not require major additional software development.

Secondary Benefits The above-mentioned **challenge of specifying the semantics of datasets is not unique to mathematics** — in all sciences, it is vital to share not only the raw data but also the description of its meaning. Mathematics is just an area in which the latter is particularly difficult; on the other hand, mathematics is also the domain, where — due to the inherent complexity of the domain — methods and practices have been developed to deal explicitly with the semantics of objects and their relations.

Because mathematical data is so rich in structure, a framework like ours that can elegantly represent all mathematical data can also represent virtually all scientific data. After all — according to none less than Galileo — “mathematics is the language in which God has written the universe”. Therefore, **the solutions for mathematical datasets developed in the *FAIRMat* project will also carry over and thus benefit other sciences**.

1.2 Relation to the Work Programme

This proposal relates to the topic “Prototyping new innovative services (INFRAEOSC-02-2019)” of the call “Implementing the European Open Science Cloud (H2020-INFRAEOSC-2018-2020)”.

1.2.1 Specific Challenge

The objectives of *FAIRMat* fit the scope of the call perfectly. There are currently a variety of highly innovative and widely used services for digital data in the mathematical sciences. These are developed by researchers, often in very agile open source communities, in response to the specific needs in their own community and therefore fit their purpose exactly.

But they currently form a patchwork of disparate and mostly ad-hoc — albeit mature and powerful — services. Therefore, the *FAIRMat* consortium brings together a representative selection of experts in the development, maintenance, and application of such services. *FAIRMat* will integrate the most powerful and most used of these services into a coherent scalable easy-to-use service offering for the mathematical sciences that can be readily deployed on the EOSC Hub.

In the mathematical sciences, a systematic collaboration between Open Science-committed and user-oriented service providers is very innovative. Our proposal is inspired by and partially driven by the results of the OpenDreamKit project (Horizon 2020, 2015–2019), which pioneered this collaboration model.

1.2.2 Scope

The EOSC Hub has so far mostly focused on *generic* Open Science services, i.e., services that can be applied uniformly to all datasets from all disciplines. While this has led to a very powerful service offering, it has gaps when it comes to the needs of specific scientific communities. This applies in particular to mathematical data, including both data from mathematics itself as well as mathematically structured data from other disciplines. **Virtually the entire research data cycle for mathematical data requires semantics-aware services**, i.e., services that are aware of and can leverage the internal structure of the datasets instead of treating the entire dataset as a whole. This includes, e.g., the generation, maintenance, application, archival, and reuse of data.

FAIRMat fills this gap. The *FAIRMat* consortium is carefully chosen to **cover the most mature systems and technologies that are currently available** for the semantics-aware processing of mathematical data (see Section 1.4.6 for technologies, their TRLs, and how they relate to the consortium). The *FAIRMat* services respond directly to the particular needs in those communities, in particular the coherent integration, systematic deployment, and the general improvement and scaling up of these technologies.

While initially driven by these needs, *FAIRMat* eventually builds more than the sum of its parts. By integrating mathematical datasets via a uniform standard (see WP2) and mathematical services through a uniform platform (see WP3), **we make them available to much larger interdisciplinary communities**. In particular, *FAIRMat* includes the development of multiple client applications (see WP3) for our services. Crucially, these are integrated with existing widely-used systems, thus making it possible for users from other disciplines and industry to discover our services and integrate them into their existing work flows.

FAIRMat is **strongly committed to providing a prototype service that can be readily integrated with the EOSC** (see WP3). To maximize our impact, we ensure that many representative and well-known mathematical datasets out there, like the ones surveyed in [Bera; Berb], will be already deployed on this prototype service (see Figure 1.3.5 and WP4).

Besides increasing the popularity of the EOSC, this will provide a well-greased pathway for other users to share their data via the EOSC. In particular, this can salvage the many large and practically used datasets that are currently generated and lost soon thereafter. The latter happens because many datasets are created in the scope of small underfunded or unfunded research projects, often by junior researchers or PhD students, who are currently forced to abandon their datasets when they change research areas or pursue a non-academic career.

1.3 Concept and Methodology

1.3.1 Overall concept and main ideas

1.3.1.1 The FAIR Principles in Mathematics The FAIR principles as laid out in, e.g., [Wil+16] are strongly inspired by scientific datasets that contain arrays or tables of simple values like numbers. In these cases, it is usually relatively easy to achieve FAIRness. But in mathematics and related sciences, the objects of interest are often much less uniform and highly structured entities, and it is important that any object in a dataset is accessible, findable, reusable, and interoperable individually. As a consequence, **the representation and modeling of mathematical data is much more difficult than anticipated in [Wil+16]**. In the sequel, we discuss the four FAIR principles and the challenges they pose for mathematical data in increasing order of difficulty.

Accessible Mathematical datasets are typically accessible in the sense of FAIR: they are made available online, include metadata, and can be retrieved via their identifier using standardized and open protocols. However, this does not allow accessing their rich internal structure. **The level of accessibility needed in practice is much harder due to the wide variety of internal structure in mathematical datasets**.

Because this functionality is so critical, many mathematical datasets are already shared in a way that assigns persistent and globally unique identifiers to each entry in the dataset or even to every subobject of each entry (e.g., OEIS [OEIS], LMFDB [LM], FindStat [BSa14], and others). But this is usually done ad hoc, identifiers are not standardized across datasets and may not be persistent, and communication protocols are dataset-specific.

Reusable Mathematical datasets are typically not reusable or very hard to reuse in the sense of FAIR. First of all, they are often shared without licenses with the implicit, but legally false assumption that putting them online makes them public domain. More critically, the associated documentation often does not cover how precisely the data was created. This documentation is usually provided in ad hoc text files or implicitly in journal papers or software source code that potential users may not be aware of and whose detailed connection to the dataset may be elusive.

The problem is that the meaning and provenance of mathematical data must usually be given in the form of complex mathematical data themselves — not just as simple metadata that can be easily annotated. And **the lack of a standard for associating complex semantics and provenance data effectively precludes or impedes most reuse in practice**.

Findable Mathematical data is typically somewhat findable in the sense of FAIR in that objects have globally unique identifiers, are connected to metadata via these identifiers, and are registered in searchable resources. This is particularly successful for bibliographic metadata (e.g. in Math Reviews, zbMATH or swMATH). However, for individual datasets, identifiers are often non-persistent, e.g., when shared on researchers' homepages.

But in any case, finding a mathematical object by its identifier or metadata is an easy problem in practice. It is much more important and difficult to find objects according to their internal structure or semantic properties. The indexing necessary for this is very difficult.

For example, consider an engineer who wants to prevent an electrical system from overheating and thus needs a tight estimate for the term $\int_a^b |V(t)I(t)|dt$ for all a, b, where V is the voltage and I the current. Search engines like Google are restricted to word-based searches of mathematical articles, which barely helps with finding mathematical objects because there are no keywords to search for. Computer algebra systems cannot help either since they to do not incorporate the necessary special knowledge. But the needed information is out there, e.g., in the form of

Theorem 17. (Hölder's Inequality)
If f and g are measurable real functions, $l, h \in \mathbb{R}$, and $p, q \in [0, \infty)$, such that $1/p + 1/q = 1$, then

$$\int_l^h |f(x)g(x)|\,dx \leq \left(\int_l^h |f(x)|^p\,dx\right)^{\frac{1}{p}} \left(\int_l^h |g(x)|^q\,dx\right)^{\frac{1}{q}} \tag{1}$$

and will even extend the calculation $\int_a^b |V(t)I(t)|dt \leq \left(\int_a^b |V(x)|^2\,dx\right)^{\frac{1}{2}} \left(\int_a^b |I(x)|^2\,dx\right)^{\frac{1}{2}}$ after the engineer chooses $p = q = 2$ (Cauchy-Schwarz inequality). Estimating the individual values of V and I is now a much simpler problem.

Admittedly, Google would have found the information by querying for "Cauchy-Schwarz Hölder", but that keyword itself was the crucial information the engineer was missing in the first place. In fact, **it is not unusual for**

mathematical datasets to be so large that determining the identifier of the sought-after object is harder than recreating the object itself.

Interoperable The FAIR principle base interoperability on describing data in a "formal, accessible, shared, and broadly applicable language for knowledge representation". But due to the semantic richness of mathematical data, **defining an appropriate language to allow for interoperability is a hard problem itself**. Therefore, existing interoperability solutions tend to be domain-specific, limited, and brittle.

For trivial examples, consider the dihedral group of order 8, which is called D_4 in SageMath but D_8 in GAP due to differing conventions in different mathematical communities (geometry vs. abstract algebra). Similarly, $0°C$ in Europe is "called" $271.3°K$ in physics. In principle, this problem can be tackled by standardizing mathematical vocabularies, but in the face of millions of defined concepts in mathematics, this has so far proved elusive. Moreover, large mathematical datasets are usually shared in highly optimized encodings (or even a hierarchy of consecutive encodings), which knowledge representation languages must capture as well to allow for data interoperability.

The proposers have developed or been involved with multiple leading candidates for such representation languages that will be integrated into a standard language by *FAIRMat*.

1.3.1.2 Kinds of mathematical data

A main idea of *FAIRMat* is a **novel categorization of mathematical data**, which allows analyzing the specific challenges to FAIR data sharing. An overview is given in Figure 1.3.1. Each kind of data has distinct strengths and weaknesses that a universal approach must take into account. Some of those strengths and weaknesses are summarized in Figure 1.3.2.

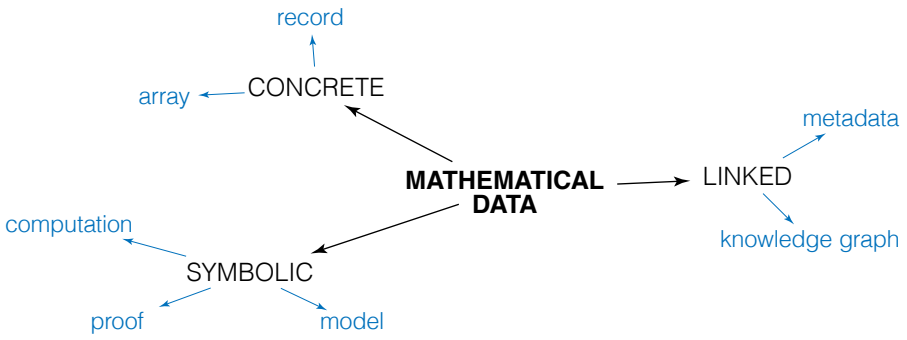

Figure 1.3.1: Kinds of mathematical data

Symbolic data consists of formal expressions such as formulas, formal proofs, programs, graphs, diagrams, etc. These are written in a variety of highly-structured formal languages specifically designed for individual domains. Because it allows for abstraction principles such as underspecification, quantification, and variable binding, symbolic data can (in contrast to the other two kinds) capture the full semantics of mathematical objects. This comes at the price of being context-sensitive: **expressions cannot be easily moved across environments, which makes *Finding*, *Reusing*, and *Interoperability* difficult**.

Working with symbolic data in mathematics can be subdivided based on the area of application into **modeling**, **deduction**, and **computation**. Each area employs a wide variety of sophisticated formal languages: modeling languages, logics, resp. programming languages.

Multiple different, often mutually non-interoperable, representation formats have been developed for symbolic data, usually growing out of small research projects and reaching different degrees of standardization, tool support, and user following. These are usually optimized for specific applications, and little cross-format sharing is possible. In response to this problematic situation, standard formats have been designed such as MathML [MML310] and OMDoc/MMT [MMTa]. The latter has been used as an interoperability format for computer algebra systems in the OpenDreamKit project and already offers comprehensive services for symbolic data such as querying. We get back to this in Section 1.4.1.

Concrete data employs representation theorems that allow encoding mathematical objects as simple data structures built from numbers, strings, lists, and records. Thus, contrary to the other two kinds of mathematical data, concrete data combines optimized storage and processing with capturing the whole semantics of the objects. But such representation theorems do not always exist because sets and functions, which are the foundation of most mathematics, are inherently hard to represent concretely. Moreover, representation theorems may be very difficult to establish and understand, and there may be multiple different representations for the same object.

In any case, *Access* is difficult because users need to know the representation theorems to understand the encoding, and this is often very complex. Therefore, **even if the representation function is documented, *Finding*, *Reuse*, and *Interoperability* are difficult and error-prone**. For example, consider the following very recent incident from (Jan. 2019): There are two encoding formats for directed graphs, both called `digraph6`: Brendan McKay's [McKb] and the one used by the GAP package Digraphs [Beu+], whose authors were unaware of McKay's format and essentially reinvented a similar one [DG]. The resulting problem has since been resolved but not without causing some misunderstandings first.

Concrete data can be subdivided into **record** data, where datasets are sets of records conforming to the same schema, and **array** data, which consists of very large, multidimensional arrays that require optimized management. Array data tends to come up in settings with large but simply-structured datasets such as simulation time series, while record data is often needed to represent complex objects, especially those from pure mathematics. Record data and querying is very well-standardized by the relational (SQL) model. However, if encodings are used, SQL can never answer queries about the semantics of the original object. The younger array data bases, which offer efficient access to contiguous — possibly lower-dimensional — sub-arrays of datasets (voxels), are less standardized, but OPenNDAP [ODAP] is becoming increasingly recognized even outside the GeoData community, where it originated.

Linked data introduces identifiers for objects and then treats them as blackboxes, only representing the identifier and not the original object. The internal structure and the semantics of the object remain unspecified except for maintaining a set of named relations and attributions for these identifiers. The named relations allow forming large networks of objects, and the attributions of concrete values provide limited information about each one. Linked data can be subdivided into **knowledge graphs** and **metadata**, e.g., as used in publication indexing services.

As linked data forms the backbone of the Semantic Web, linked data formats are very well-standardized: data formats come as RDF [SR14], the relations and attributes are expressed as ontologies in OWL2 [OWL09], and RDF-based databases (also called triplestores) can be queried via SPARQL [HS13]. For example, services like DBPedia [Leh+13] and Yago [Hof+13] crawl various aspects of Wikipedia to extract linked data collections and provide SPARQL endpoints. The WikiData database [WD] collects such linked data and uses them to answer queries about the objects.

Thus, contrary to the other two kinds, linked data has very good FAIR-readiness, in particular allowing for URI-based *Access*, efficient *Finding* via query languages, and URI-mediated *Reuse* and *Interoperability*. However, this **FAIR-readiness comes at the price of not capturing the complete semantics of the objects so that *Access* and *Finding* are limited and *Interoperability* and *Reuse* are subject to misinterpretation**.

Kind of data	Symbolic	Concrete	Linked
Allows recovering the represented object	+	+	−
Applicable to all objects	+	−	+
Easy to process	−	+	+

Figure 1.3.2: Advantages of different kinds of data

1.3.1.3 Semantics-aware Open Data and Deep FAIRness Concrete and linked data can be easily processed and shared using standardized formats such as CSV (comma-separated value lists) or RDF [OWL09]. But in doing so, the semantics of the original data is not part of the shared resource: in concrete data, the semantics requires knowing the encoding function; and in linked data, almost the entire semantics is abstracted away. For datasets with very simple semantics, this can be remedied by attaching informal labels (e.g., column heads), metadata, or free-text documentation. But this is not sufficient for datasets in mathematics and related scientific disciplines where the semantics is itself very complex.

For example, an object's semantic type (e.g., "polynomial with integer coefficients") is typically very different from the type as which it is encoded and shared (e.g., "list of integers"). The latter allows reconstructing the original, but only if its type and encoding function (e.g., "the entries in the list are the coefficients in order of decreasing degree") are known. Already for polynomials, the subtleties make this a problem in practice, e.g., consider different coefficient orders, sparse vs. dense encodings, or multivariate polynomials. Even worse, it is already a problem for seemingly trivial cases like integers: for example, the various datasets in the LMFDB use at least 3 different encodings for integers (because the trivial encoding of using the CPU's built-in integers does not work because the involved numbers are too big). But mathematicians routinely use much more complex objects like graphs, surfaces,

Service	Shallow	Deep
Identification	DOI for a dataset	DOIs for each entry
Provenance	who created the dataset?	how was each entry computed?
Validation	is this valid XML?	does this XML represent a set of polynomials?
Access	download a dataset	download a specific fragment
Finding	find a dataset	find entries with certain properties
Reuse	impractical without accessible semantics	
Interoperability	impossible without accessible semantics	

Figure 1.3.3: Examples of shallow and deep FAIR services

or algebraic structures.

We speak of **accessible semantics** if data has metadata annotations that allow recovering the exact semantics of the data. Notably, in mathematics, this semantics metadata is very complex, usually symbolic, mathematical data itself that cannot be easily annotated. But **without knowing the semantics, mathematical datasets only allow FAIR services that operate on the dataset as a whole**, which we call **shallow** FAIR services. But it is much more important to users to have **deep** services, i.e., services that process individual entries of the dataset. Figure 1.3.3 gives some examples of the contrast between shallow and deep services. Deep services are only possible if the service can access and understand the semantics of the dataset. While shallow FAIR services are easy to build, deep services are essentially non-existent except when built ad hoc for individual datasets. Figure 1.3.4 gives an overview.

Data	Findable	Accessible	Interoperable	Reusable
Symbolic	Hard	Easy	Hard	Hard
Concrete	Impossible without access to the encoding function			
Linked	Easy but only applicable to the small fragment of the semantics that is exposed			

Figure 1.3.4: Deep FAIR readiness of mathematical data

Note that **the advantages of deep services are not limited to mathematics at all**. For example, in 2016 [ZEE16], researchers found widespread errors in papers in genomics journals with supplementary Microsoft Excel gene lists. About 20% of them contain erroneous gene name because the software misinterpreted string-encoded genes as months. In engineering, encoding mistakes can quickly become safety-critical, i.e., if a dataset of numbers is shared without their physical units, precision, and measurement type. With accessible semantics, datasets can be validated automatically against their semantic type to avoid errors such as falsely interpreting a measurement in inch as a measurement in meters, a gene name as a month, or a column-vector matrix as a row-vector matrix.

A Data Representation Standard and Implementation Framework for Deep FAIR Services
We can now circle back to our objectives and state them more concisely: **the objective of *FAIRMat* is building Deep FAIR services for mathematics and related sciences**. The central idea is to integrate the existing standards for different kinds of data into a coherent representation standard for mathematical data that systematically makes the semantics accessible. This enables (i) prototyping universally applicable Deep FAIR services that improve on the existing ad hoc or limited solutions and (ii) making a wide variety of existing datasets available via a central platform.

1.3.2 Trans-disciplinary Considerations Use of EOSC stakeholder knowledge

Even though the *FAIRMat* consortium has only seven members, it brings together partners representing all the major stakeholders in mathematical data:

- **Dataset authors** are represented by the sites **FAU** (OAF) for symbolic, **FIZ** (zbMATH) for linked, and **UL** (graph datasets), and **CHA** (LMFDB) for concrete data.
- **Mathematicians as users** are represented by the sites **UPSud**, **CHA** and **UL** — in fact, the **UPSud** PI Nicolas Thiery only initiated the OpenDreamKit project in order to improve the tools so that he could better

pursue his actual research. Similarly, the PIs Primož Potočnik at **UL** and Stefan Lemurell at **CHA** need services both as providers and users of their datasets.
- **Users from other sciences** are represented by **CAE** and **EMS** (via the PIs' personal research) for the modeling of cyber-physical systems.
- **Service providers** are represented by the sites **FAU** (MathHub and MathWebSearch) and **CAE** (Emmo) for symbolic data, **FIZ** (zbMATH and swMATH) for linked data, and **CHA** (LMFDB) and **UL** (DisreteZOO) for concrete data.
- **Mathematical institutions** are represented by the sites **EMS** (European Mathematical Society) in general and for mathematical knowledge bases by **FAU** and **FIZ** (the PIs are members of the International Mathematical Knowledge Trust IMKT).

This allows optimally leveraging stakeholder knowledge, networks, and user communities.

1.3.3 Linked activities: existing datasets and services

The sites **FAU** and **UL** are collaborating on a living survey of concrete record datasets [Bera], which has served as a market study for the *FAIRMat* project. This study was very well received in the mathematical community, and word-of-mouth even led to the discovery of several orphaned datasets. To maximize impact and advertise FAIR sharing, we commit to already integrating a significant representative collection of datasets during the project lifetime. The concrete selection presented in this section is driven by the size and importance of the datasets, the sizes of the involved user groups, and the goal of reaching a diverse set of users. See Section 1.3.4.2 for details.

Concretely, *FAIRMat* will incorporate the datasets from Figure 1.3.5 and link to the external infrastructures summarized in Figure 1.3.6. Note that the technologies used internally in *FAIRMat* are listed and discussed separately in Section 1.4.6.

1.3.4 Overall Methodology

1.3.4.1 Open data framework

We will develop a framework for representing mathematical datasets using symbolic, concrete, and linked data with accessible semantics.

Because the distinct advantages of the three kinds of data are very difficult to combine, we allow each kind and integrate all kinds into a coherent whole. This will be based on established open formats:
- For symbolic data, we use the OMDoc representation language. It provides uniform encodings of symbolic data in a single standardized concrete language.
- For concrete data, we use record and array representations based on standard formats such as JSON. To make its semantics accessible, we use the codec framework developed in the OpenDreamKit project (see Section 1.4.1.2 for details). It provides a systematic link between symbolic specification of mathematical objects and their encoded representation.
- For linked data, we use RDF and OWL. To make the semantics of linked data identifiers accessible, we use MMT URIs that allow symbolic and linked data to share the same identifiers.

All datasets and all objects in them will have URIs via which they become accessible. We will also define a metadata standard that allows for tracking provenance, version, and license of mathematical datasets and their entries. These URIs and the associated metadata form a linked dataset themselves, which is also stored in the framework.

Finally, to ensure the sustainability of the *FAIRMat* standard, we will submit it for ISO standardization by the end of the project duration. These research efforts are detailed in WP2 as well as (for the ISO standard) in WP1.

1.3.4.2 Pilot datasets

We integrate a representative selection of major datasets from different areas of mathematics into our infrastructure.

We have carefully put together our consortium such that these communities are represented by partners, i.e., for all pilot datasets there are partners who are the maintainers themselves or have close ties to them.

These pilot datasets were chosen for multiple purposes:
- We respond to the existing needs of the communities maintaining the datasets.

Task	Site	PI contact	Dataset	Maintainer(s)		Description
Symbolic Data						
T4.1	**FAU**	Florian Rabe	Theorem Prover Libraries	[OAF]	**FAU**	≈ 5 proof libraries $\approx 10^5$ theorems ≈ 50 GB
T4.2	**CAE**	Peter Harman	Modelica libraries: > 10 official > 100 open-source ≈ 50 commercial	[MOb]	**CAE**	> 5.000 classes in the Standard Library, industrial models can reach .5M equations
Record Data						
T4.3	**FAU**	Michael Kohlhase	Integer Sequences	[OEIS]	OEIS Foundation	≈ 330K sequences ≈ 1 TB
		Michael Kohlhase	Sequence Identities	[KDH]	**FAU**	$\approx .3$M sequence identities ≈ 2.5 TB
	UL	Katja Berčič Primož Potočnik	Highly symmetric graphs, maps, and polytopes	[COa] [AP] [AAP] [COb] [TG] [EET]	Marston Conder Michael Hartley Dimitri Leemans Primož Potočnik Gordon Royle Stephen Wilson	≈ 30 datasets $\approx 2 \cdot 10^6$ objects ≈ 1 TB
		Katja Berčič	Finite lattices	[FL] [AG] [UL]	Jukka Kohonen Dimitri Leemans Martin Malandro	7 datasets $\approx 17 \cdot 10^9$ objects ≈ 1.5 TB
	UPSud	Nicolas Thiery	Combinatorial statistics and maps	[BSa14]	FindStat	≈ 1.500 objects
		Nicolas Thiery	SageMath databases	[SDB]	SageMath	12 datasets
T4.4	**CHA**	Stefan Lemurell	L-functions, Modular Forms, Elliptic Curves	[LM]	The LMFDB Collaboration	≈ 80 datasets $\approx 10^9$ objects ≈ 1 TB
Linked Data						
T4.5	**FIZ**	Olaf Teschke	zbMATH	[ZBM]	**FIZ**	$\approx$ 4M publication records with semantic data ≈ 30M reference data > 1M disambig. authors $\approx 2,7$M persistent links to full texts, thereof ≈ 1M OA
		Olaf Teschke	swMATH	[SWM]	**FIZ**	$\approx$ 25K software records with $>$ 300K links to $>$ 180K publications
	EMS FIZ	Olaf Teschke	EuDML	[EUD]	**FIZ**	≈ 260K open full-text publications
T4.6	**FIZ**	Moritz Schubotz	Wikidata	[WD]	Wikimedia Foundation	34 GB linked data, thereof about 4K formula entities, interlinked, e.g., with named theorems, persons, and/or publications

Figure 1.3.5: Summary of datasets that will feed into the project *FAIRMat*

Infrastructure	Maintainer	Description	PI contact
EOSC/EUDat	EU	European Open Science Cloud	Florian Rabe (**FAU**)
GitHub	GitHub, Inc.	data repositories, e.g. for Modelica	Peter Harman (**CAE**)
RADAR	**FIZ**	Research Data Repository	Matthias Razum (**FIZ**)
Wikimedia	Wikimedia Foundation	Mathematical data	Moritz Schubotz (**FIZ**)
DLMF	NIST	Digital Library of Mathematical Functions	Michael Kohlhase (**FAU**)
arXiv.org	Cornell U.	1.5M Preprints in Science	Michael Kohlhase (**FAU**)

Figure 1.3.6: Summary of Infrastructures that feed into *FAIRMat*

- They serve as test cases and evaluation case studies.
- Their detailed documentation provides blueprints and tutorials that will be made available for other communities.
- They ensure high visibility of *FAIRMat* and thus contributes to long-term adoption.
- The involved communities provide a starting point for outreach workshops.

Figure 1.3.5 gives an overview of the datasets we have chosen. It also indicates (i) how formidable the challenge is given the number and sizes of the datasets and (ii) how well the *FAIRMat* consortium matches the challenge, with each partner being an expert on one of these datasets. These research efforts are detailed in WP4.

1.3.4.3 Service prototypes and their integration into the EOSC Hub

The core service prototyped by *FAIRMat* is the semantics-aware FAIR data sharing infrastructure that allows the uniform integration and interoperation of mathematical services across all datasets.

This service will be deployed on a major server or small cluster of servers that are funded by *FAIRMat*. A reference instance of these services will be maintained by the coordinating site **FAU**, but all hardware and software will be designed such that the services can be easily ported or replicated by other providers such as the **FIZ** site. These servers can be maintained for some time beyond the *FAIRMat* life time.

In addition several advanced innovative services will be realized that are enabled by the accessible semantics of the datasets. These are deep citability, versioning, and provenance tracking; validation, browsing and visualization; as well as computation including the integration with existing computation systems (see WP3 for details).

To ensure that our service prototypes can be integrated into the EOSC Hub at the completion of *FAIRMat*, we undertake some efforts already during the *FAIRMat* lifetime. These include in particular the specification of hardware and software interfaces, legal issues, and accessibility requirements (e.g., those set out in projects funded under the EINFRA-12-2017 topic). These research efforts are detailed in WP3.

1.3.4.4 Outreach and Adoption in Different User Communities

We ensure the scalability of our results by providing our services to user communities from different disciplines during the *FAIRMat* lifetime.

We will start with outreach activities immediately at the start of the project and gradually ramp them up. This will also aid with raising the awareness of FAIR concepts in the mathematical community.

Concretely, we adopt a multi-pronged approach. Firstly, we engage existing dataset maintainers and support them in sharing their datasets via the existing EOSC services. If necessary, we offer doing this for them in order to decrease the necessary efforts on their side. This will quickly and for the first time create a collection of a few dozen datasets from all kinds of mathematical fields accessible in a single place. These datasets will create a critical mass that we can advertise in informal media, formal publications, and at workshops. Most importantly, we will make the official announcement of this dataset collection at the European Congress of Mathematicians in 2020. Because this is already a major improvement on the current disparate and ad-hoc nature of dataset sharing in mathematics, this will generate substantial visibility of and interest in *FAIRMat*.

Secondly, the above initial collection of datasets will initially not be reusable or interoperable, let alone searchable, because the datasets will not yet conform to a common standard. To develop requirements for, advertise, and collect feedback on our standard and the enabled advanced services, we will organize two extended workshop

events, which we dub “Summer of Math Data”. They will feature a series of partially overlapping research visits of individual dataset providers, anchored by short workshops and conferences open to all mathematicians.

Finally, we will organize topical workshops at mathematical software conferences like the ICMS and CICM. And in the summer of 2022 we will organize a major workshop at the 4-annual International Congress of Mathematicians (ICM), where we will officially release the final results of *FAIRMat*. These efforts are detailed in WP5.

1.3.5 Gender Dimension

Mathematical data is inherently sex and gender-neutral, and this extends to all innovations and services of *FAIRMat*. Examples and use cases in the generated documentations will be chosen in sex/gender neutral ways.

Linked data can facilitate research and impact policies on gender issues. Already, zbMATH data plays an essential role in the Gender Gap in Science Project [GGS], led by the International Mathematical Union (IMU) and the International Union of Pure and Applied Chemistry (IUPAC) and supported by other major global scientific organizations, which aims to evaluate different career paths in relation to gender, identify obstacles, and derive recommendations. Making this information available as FAIR data could not only spur further research and result in appropriate policies, but also facilitate the creation of much more granular information. E.g., the project has already identified a significant impact of the geographical and subject data on gender-specific career paths. Further relevant information on these issues would be intrinsically generated by interlinking the various data within the project and made available as a unique new FAIR source.

All partners will follow inclusive practices in recruiting staff for this project, in inviting the community to our workshops and outreach events and in choosing users to evaluate prototypes.

1.4 Ambition

1.4.1 State of the Art

For most mathematicians, and related researchers and users, the 2019 state of the art in data sharing often still consists of posting their dataset online, e.g., as a CSV-encoded text file or uploading it to a database with a web interface. This leads to a patchwork of individual databases and service offerings that do not meet the FAIR criteria beyond the most basic level of accessibility. They are not interlinked, their semantics may not be accessible, and chosen encodings are usually incompatible across datasets. And even accessibility is limited by both limited documentation and API access restrictions.

Cut-and-paste or email-based exchange of datasets are very common, and even if central storage services like GitHub are used, the datasets may not be well-documented enough to allow for unsupervised reuse by other researchers or even the same researcher at a later time. In a typical reuse scenario, a user would find a dataset via general purpose internet search engine or word of mouth, access it through a web interface, manually figure out the encoding of the result, and then write an ad-hoc piece of code that transcodes the result into the format needed for further computation. This piece of code is error-prone, might not be rerunnable automatically, and tends to decay quickly even if shared.

For an instructive example of the state of the art, consider that the Sage [Sage] system routinely bundles a copy of the LMFDB dataset of elliptic curves [LM] because a live integration is too brittle and inefficient; even worse those copies are not even generated automatically from the LMFDB — instead John Cremona, the maintainer of the dataset, manually runs a script to produce the copy from his own raw data every time he updates the LMFDB dataset. One inspiration of the present proposal is a service we developed [D6.518] in the OpenDreamKit project, which allows Sage users to directly query elliptic curves from the LMFDB. This systematic approach also revealed inconsistencies in the LMFDB encodings and helped the LMFDB maintainers address them.

Note that the above description is not meant to disparage previous efforts. In fact, substantial knowledge and data sharing services have been built, and in the sequel we discuss some of the most pertinent. But it is helpful to understand why the problem is so difficult; therefore, in Section 1.4.2, we discuss the challenges that make Open Data solutions harder for mathematics than for other sciences.

1.4.1.1 Computation Systems Multiple computer algebra systems have been developed that include major sharing of software and data. Mathematica [WM] and SageMath [Sage] are widely-used systems that are interesting to discuss as representative examples.

Mathematica is a closed source commercial system by Wolfram Research that aims at a uniform coherent interface to all mathematics. The Wolfram language and the Mathematica notebook concept have influenced many other systems. It also features a large collection of datasets from mathematics, science, and general statistics about the world that are accessible to computation. Parts of the systems are free (of charge), e.g., the Wolfram—Alpha web interface and a large collection of Mathematica packages that users contributed independently of Wolfram. But the core system is proprietary and owned by a US company, which makes it unsuitable as a platform for FAIR mathematics in the EOSC.

SageMath is developed under the GPL by a large international user community led by William Stein, the system's original developer. It was specifically created as an Open Source alternative to dominant computer algebra systems like Mathematica. SageMath pulls in as many software packages as possible and uses Python to build an integration layer at which these unrelated packages can communicate. SageMath has become extremely popular among mathematicians, especially those active in the Open Source community and was a central component in the OpenDreamKit project. Due to its size and integrative nature, it has become somewhat of a cross–programming language packaging and distribution platform, via which researchers can disseminate specialized Open Source computation libraries. It integrates a number of important mathematical datasets such as copies of some LMFDB datasets or GAP's small groups library.

Note that there is a blurry line between computation and data in mathematics — and any science where data is computed as opposed to measured: instead of storing data, we can always recompute it on demand; dually, computation can be replaced by tabulation of results. Which one is better, depends on the trade-off between time and space, i.e., computing and storage costs. And this trade-off in turn changes constantly according to hardware power and costs as well as community size and usage patterns.

1.4.1.2 Semantic Interoperability The SageMath integration layer described above is largely non-semantic in the sense that it relies on custom "glue code" in Python that is unverified and can be broken by any update of one of the integrated systems or datasets. Moreover, it is specific to SageMath and cannot be reused in other systems.

To have a more semantic, maintainable, and flexible interoperability platform, the OpenDreamKit project developed the Math-in-the-Middle (MitM) Framework [D6.518].

There, all systems (boxes A-H in Figure 1.4.1) export an interface specification (circles a-h) as symbolic data, whose concepts are aligned with a knowledge graph representing abstract — i.e. system-independent — mathematical knowledge (the MitM ontology in the center of 1.4.1). Systems are connected by a MitM mediator system, which reads the interface specifications, the MitM ontology, and the alignments and translates between the OpenMath-based system dialects. This mediator is built on top of and inherits most of the functionality from the MMT API [MMTc], an open knowledge management system for symbolic data developed at **FAU**.

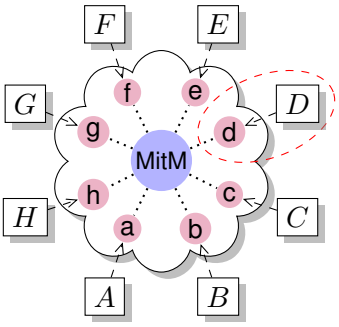

Figure 1.4.1: Math-in-the-Middle Interoperability Framework

Figure 1.4.2 shows a workflow and mathematical use case that makes use of the flexible delegation of sub-problems to external systems, which is not possible in SageMath alone. Additionally, this MitM workflows can be set up just as easily with other systems as the master, e.g., where a user working in the computer algebra system GAP [GAP] delegates to SageMath and the number theory library PARI/GP [PG]; see [Koh+17] for details.

In addition to computation systems and as a major inspiration for the *FAIRMat* proposal, MitM also allows integrating mathematical *databases*. Here the MitM mediator directly connects to the database-level API and automatically decodes the encoded concrete mathematical data into mathematical objects via semantic annotations to the database tables. The latter uses the mathematical schema technology developed in OpenDreamKit to specify the mathematical structure and encodings of concrete data.

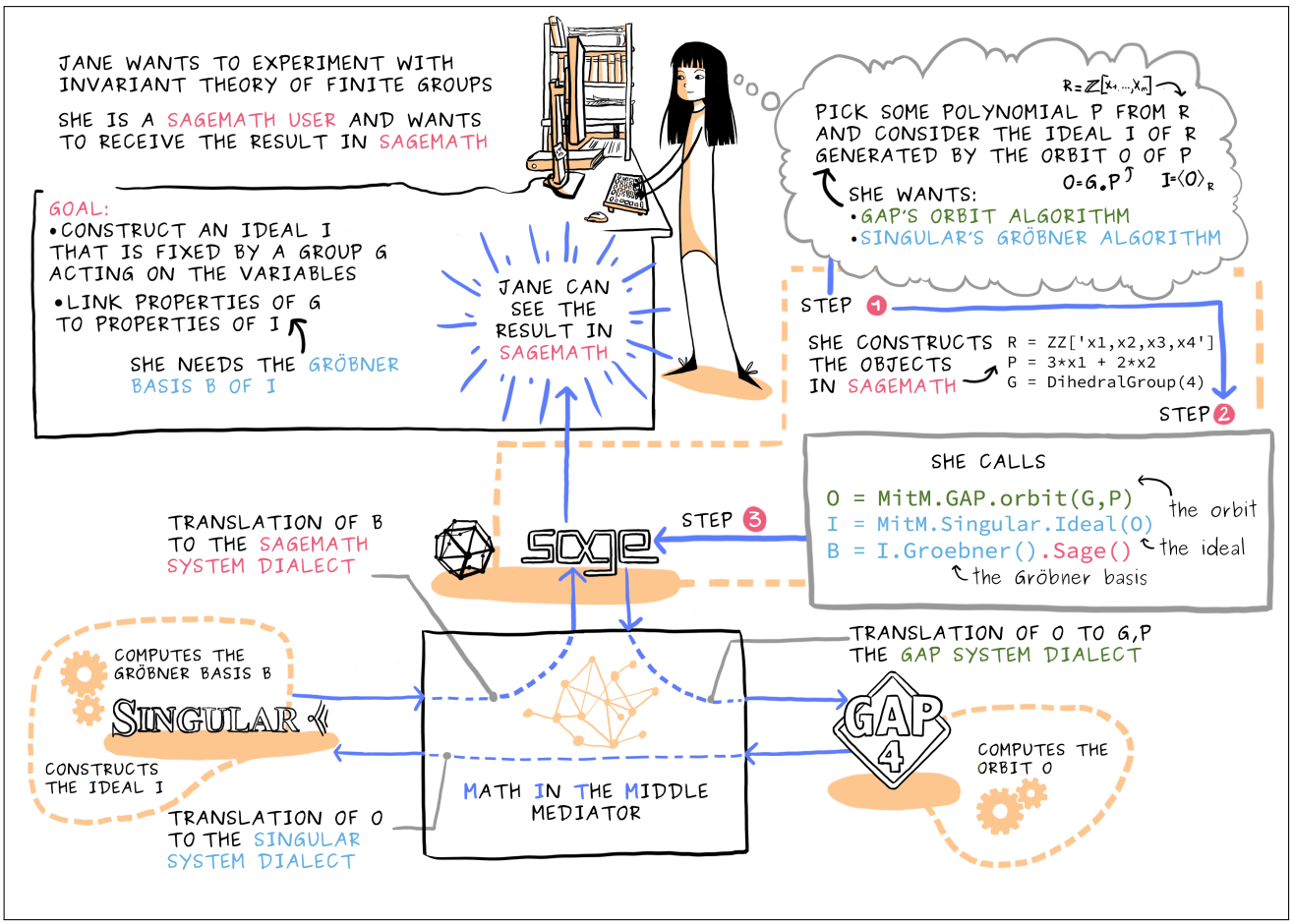

Figure 1.4.2: OpenDreamKit: A MitM-based Workflow

1.4.1.3 Symbolic Representation Languages Symbolic data requires very sophisticated formal languages, such as type theories, programming languages, and logics to represent. This has led to a major fragmentation into dozens of incompatible languages, each one with a major library.

While concrete and linked datasets often have simple structure but then large size (i.e., a lot of records all conforming to the same schema), symbolic datasets tend to work with very different individual statements (e.g., many theorems for very different formulas). In the area of deduction, Coq [Coq] and HOL Light [Har96] are among the most well-known ones because they were used in flagship projects — the proofs of the Feit-Thompson theorem resp. the Kepler conjecture. In the area of modeling cyber-physical systems, Modelica [MOb] has been very influential. The success of Modelica is largely due to the ease of model reuse, which in turn is due to technical choices in the language design, such as being object-oriented and equation-based; and due to the concept of packages, or libraries. This package concept has resulted in an ecosystem of library development, both open- and closed-source, and through academia and industry. Dependencies between libraries, and between individual models in libraries, can be complex.

The Modelica language is an open-standard, developed by the Modelica Association, and multiple different implementations exist, owed in part to its commercial relevance. Different implementations can have different mechanisms for sharing or locating models or libraries.

All of these systems tend to have open licenses for both the source code and the library. While FAIR sharing within each system is quite easy (except that search is notoriously difficult), the sharing across languages is extremely difficult and often prohibitively expensive in practice. MathML [MML310] and OMDoc [Koh06] are XML-based standardized representation languages that were developed in order to overcome this segregation into disconnected islands of FAIRness. MathML in particular was sanctioned by the World Wide Web Consortium (W3C) and is partially included into the HTML5 standard. It has two sub-languages: presentation MathML for the layout in browsers etc. and content MathML for the meaning of formulas; the latter is based on the OpenMath standard [Bus+04], which represents symbolic data in the form of operator trees using function application, variables, binding, and symbols. However, practical adoption has been limited due to the high cost and lack of incentives for researchers to make their symbolic data interoperable across language boundaries.

1.4.1.4 Semantic Web and Knowledge Graphs Data frameworks based on Linked data have been used in many disciplines to represent large datasets of both general objects (e.g., SUMO [SUMO]) and domain-specific ones (e.g., OBI [OBI] for biology or SNOMED CT [SCT] for medicine). These datasets area usually written in the form of ontology languages such as OWL2 [OWL09] or RDF [RDF], which can be easily written manually or generated automatically from other datasets.

Due to its simplicity and universal applicability, this has become a very successful interface layer at which datasets can be shared and related. However, these advantages come at a price: atomic objects are treated as black boxes whose internal structure and semantics are abstracted away. That limits linked data to a small portion of the data, for which it is excellently suited.

These limitations are particularly severe in mathematical sciences, where organizing large networks of concepts is the easy part and most effort is spent on the formal semantics of the individual objects. Consequently, FAIR mathematics must complement these simple universal solutions with more complex math-specific ones.

1.4.1.5 Databases for Concrete Data A wide variety of concrete datasets has been developed. Typically, each one is maintained in a separate ad hoc database (like a text file with one entry per line) or SQL databases. A living survey of such databases can be found in [Berb; Bera]. At the time of writing, it contains about a hundred datasets, ranging from small (up to 100 objects) to very large ($\approx 17 \cdot 10^9$ objects). Similarly varied is the authorship: from one author producing several smaller datasets (Conder [COa], Potočnik [COb], McKay [McKa], Royle [Roy], Wanless [Wan]), to FindStat [BSa14] (69 contributors), LMFDB [LM] (100 contributors) and the OEIS [OEIS] with thousands of contributors.

The oldest such datasest is the online encyclopedia of integer sequences. It is notable for collecting not only mathematical objects but also symbolic data for their semantics (such as defining equations and generating functions).

LMFDB is an exception in that it is a collection of mathematically related databases; these are conceptually tied together by Langland's program [Ber03] and at a systemlevel by a set of mature generic tools and a web-based user interface that support for accessing and searching arbitrary datasets. The LMFDB has been a successful collaboration thanks to over ten years of continues funding used wisely for regular workshops attracting many

mathematicians, visionary leadership and important technical support from computer scientists. Unfortunately, the LMFDB data is today only available through a web service that lack many of the features of a deep FAIR service.

All these datasets would strongly benefit from being stored in a systematic FAIR infrastructure. But here shallow FAIR services are the easy part, so easy that most researchers barely see the benefit over ad hoc sharing via, e.g., their personal websites. Deep FAIR services are much more difficult, and are either non-existent or are supplied in a dataset-specific way.

1.4.1.6 General Data Sharing Infrastructures Several hardware and software infrastructures are available that support researchers in data sharing. It is particularly instructive to contrast the EOSC and related infrastructures such as EUDat with GitHub.

The EOSC family of services has been developed top-down based on large national and international research grants. Express goals were enabling, encouraging, and potentially requiring the FAIR sharing of research data. These services are complemented by national infrastructures like RADAR in Germany.

The institutional support allows building the large-scale long-term infrastructure that can otherwise hardly be accomplished. Moreover, the political backing allows introducing standards, which are notoriously difficult to develop in the mathematical sciences because individual research communities rarely become strong enough to "corner the market" and enforce a standard.

But the EOSC has so far received little to no attention among mathematicians. In fact, we could find only one mathematical dataset that has ever been published: Jukka Kohonen, who supports *FAIRMat* and has agreed to serve in the advisory board, published his collection of lattices. Incidentally, the size of this collection (1 TB in optimized, uncompressed representation) underlines the enormous challenge of sharing mathematical datasets. Similarly, the RADAR infrastructure [Kra+16], which offers similar services, has not been used for any mathematical dataset. By and large, mathematicians are unaware of these general data sharing infrastructures or do not gain substantially from them. Notably, even though these infrastructures offer services that are valued by mathematicians like citable identifiers (PIDs or DOIs) or persistent hosting (rather than a work group server, which usually goes away unexpectedly, e.g., at retirements), they offer *shallow* services only.

GitHub, in contrast, has emerged as a grassroots quasi-standard for sharing datasets. Indeed, a great majority of scientific software and datasets is nowadays distributed via GitHub or similar platforms. For example, the Modelica community introduced the convention of sharing libraries through GitHub in a uniform way so that GitHub combined with a crawler can be used as a decentralized package manager.

From a FAIR perspective, GitHub provides only very basic services, essentially shallow A, R, and F. Its main asset is its ease of use both for data providers and users: it is very easy to create, update, and fork datasets, give feedback on and discuss them, or use third-party tools on them. Sharing data on GitHub yields the near certainty that any potential user can use it without additional documentation, advertisement, or tool support. Of course, neither GitHub's data model (which is biased towards software projects rather than datasets) nor its organizational model (a US company providing a commercial service) is directly applicable to the EOSC. But it is important to appreciate its community penetration and its ability to dynamically adapt to user needs, which is particularly valuable to the often informally organized small communities in research projects.

1.4.2 Open data challenges specific to mathematics

Mathematics faces a number of specific challenges that make Open Data arguably harder or at least differently hard than for other sciences.

Institutional and Financial Challenge The mathematical community is mainly made up of small research groups. There are very few large research teams as are common in engineering and experimental sciences. Large collaborations (e.g., the Classification of Finite Simple Groups or the Polymath project) are driven by individuals and do not have a permanent institutional backing. Therefore, most datasets are collected and most services are provided by individual mathematicians or small communities and are not archived systematically.

This is connected to a financial challenge: even top researchers in mathematics have little research funding that can be devoted to hosting large databases and services. That precludes them from employing the computing resources necessary to host TRL 8 services, which shows the need for and potential of an EOSC-like initiative. But so far only one mathematical dataset has been published via the EOSC infrastructure — Jukka Kohonen's

collection of lattices. In conversations with him and others, we learned that this is primarily due to the lack of awareness of and the lack of semantics-aware services provided by the EOSC infrastructure.

Cultural challenge Mathematics is traditionally performed in journal articles. Even though there is a growing mathematical Open Source community, mathematicians receive little reputation and career benefit from maintaining services and sharing data. Consequently, data sharing is often only an afterthought, and mathematicians publish datasets at whatever site makes sharing easiest. Sometimes datasets are not even published at all. For example, Kohonen's dataset includes one more object than a dataset computed by colleagues in previous work. But due to the lack of services that make accessibility and reusability easy, it was impossible to determine which of the two datasets is in error.

Mathematicians also require very high standards for the reliability of the data. They expect every listed object to be 100% correct and might be wary of using data unless it has a reputable source (provenance).

This shows the need for a systematic FAIR culture in mathematics, and our experiences show mathematicians are open to this new culture. But they need a project like *FAIRMat* that provides them with a single, highly user-oriented, and widely visible service that provides dataset authors with as much added value while requiring as little added work as possible.

Technological challenge Even though most mathematicians are enthusiastic about Open Data, it is in most cases not in their interest or expertise to maintain databases and services. Even when it is, the technological difficulties often make a mathematician's time too valuable to spend on data sharing. Moreover, and contrary to other disciplines, there are many different software suites that mathematicians use to interact with datasets, even within the same area of mathematics and for a single dataset.

Therefore, mathematicians need a standard format for datasets supported by strongly user-oriented services that lets them share their data as widely, easily, and seamlessly as possible.

Theoretical challenge The rich structure and semantics of mathematics makes it impractical to apply standard techniques like SQL or RDF to mathematical objects directly. Before mathematical data can be stored (ultimately as strings of bits), it must be converted into a form that is much further from the meaning of the data as perceived by a mathematician than is usual in other sciences. To find such a representation in the first place, many related objects and conventions must be considered and most interesting mathematical objects have multiple representations. Often deep mathematical theorems must be established for each such representation.

Therefore, the standard format for datasets must be tightly integrated with symbolic representations of the objects and rigorously connect these objects to their encodings. Only then is it possible to build the semantics-aware services envisioned by *FAIRMat* that will attract mathematicians to the EOSC infrastructure.

1.4.3 Advance over state of the art

Our ambitious proposal is to provide a **comprehensive semantics-aware FAIR** Open Data solution for mathematics. The *FAIRMat* service offering will be the first of its kind in each of the following respects.

FAIR As a rule, mathematicians strongly support the Open Science movement and happily make their datasets public. This is accompanied by a vibrant and growing community of Open Source software for computational mathematics. However, today most mathematical databases are shared in an ad hoc manner that makes FAIR sharing hard to impossible. Systematic data collection and archival initiatives are limited in scope and suffer from a lack of interlinking of digital artefacts across platforms. In effect, FAIR mathematics, while widely welcomed, is effectively non-existent today. A similar argument applies to related sciences to the extent that they make heavy use of mathematical data such as the mathematical modeling of physical systems.

FAIRMat will deliver the first FAIR Open Data solution for mathematical datasets. It will provide a central service for mathematicians to make their datasets accessible. It will allow users to browse, search, retrieve, and compute with these datasets. And it will support system interoperability via automated import/export of datasets, including the transcoding from the original into the desired format.

Semantics-Aware Generally, reusing shared data requires that the reuser be able to understand the semantics of the data [FAIR18, Rec. 7]. This is particularly difficult for system interoperability where the semantics must not only be evident but must itself be accessible for automated processing [FAIR18, Rec. 8], and it is particularly critical where data is used in safety-critical systems. While this problem exists for all data, it is particularly challenging for mathematical data and similar data in related disciplines, where the semantics is very difficult to specify. Therefore, today there are virtually no mathematical datasets whose semantics is itself accessible.

FAIRMat will deliver the first Open Data solution that can *understand* mathematical data. This is essential to retain mathematical rigor in seamless Open Data collaboration where data provider and data user will often not interact with each other, e.g., data passed between systems must be translated according to its mathematical meaning, not just its textual presentation.

Understanding is also critical to realize advanced Open Data services. Without it, service providers can only offer:

- blackbox services, which treat a dataset as an atomic unit and only store and distribute it as a whole,
- graybox services, which inspect a dataset but only to a certain degree, e.g., string-based search, metadata analysis, or interlinking of fragments.

FAIRMat will allow whitebox services, which exploit the semantics of the data they act on. This will allow *FAIRMat* to build several services that have not been realized at TRL 8 before, such as

- validating a dataset against a mathematical specification,
- searching for an object based on its mathematical properties,
- translating objects from one concrete representation to another.

Comprehensive The best existing solutions for mathematical data focus on one kind of data. For example, LMFDB handles only concrete data from one narrow area of mathematics. zbMATH and swMATH handle mostly linked data for publication metadata, community reviews, software information, and symbolic data for formulas occurring in publications. Wikidata stores mostly linked data and object visualizations. But mathematics and research in general thrives on transferring results across disciplines and across environments.

FAIRMat will deliver the first Open Data solution that supports all three kinds of data — symbolic, encoded, and linked data.

Moreover, it will allow the fine-granular interlinking between all digital artefacts, overcoming the complete lack of systematic interlinking across heterogeneous digital artefacts. In particular, it will support the scalable interlinking of heterogeneous datasets, which is virtually non-existent today., e.g.,

- interlinking a database record and the code that produced it,
- interlinking a piece of code and the publication that describes the implemented algorithm,
- interlinking the publication with the formal description of its theorems.

1.4.4 Ambitiousness

Any one of the three aspects described in Section 1.4.3 constitutes an ambitions goal. But Open Data for mathematics as envisioned by the EOSC requires tackling all three at once. This comprehensive view on mathematical data constitutes a very ambitious challenge for standards development.

1.4.5 Innovation potential

As explained in Section 1.4.3, *FAIRMat* is in itself innovative, providing the first large scale replication of Open Data services for mathematics. The objective of establishing a single, coherent data framework standard and platform for mathematical data is groundbreaking.

Several of the methods employed will be innovations. While the individual services and databases already exist, *FAIRMat* will employ novel concepts to integrate them. These are in particular the data representation framework and the idea of mathematical schemas, described in Section 1.3.4. The *FAIRMat* project will focus on developing these for mathematical data, but similar ideas and principles apply to other sciences as well, and by providing a blueprint for deep FAIRness, the *FAIRMat* will act as a catalyst for innovation outside the project.

The coherent integration of the various existing services on a uniform deep FAIR platform constitutes a novel service that is a significant improvement on the current state.

Finally — while not directly targeted in the *FAIRMat* project — the hosting of mathematical datasets in a uniform semantic framework will make them very attractive for applying Machine Learning (ML) methods, e.g., to derive

new mathematical conjectures that can then be attacked by conventional methods. Such methods of *experimental mathematics* are already becoming popular in mathematics wherever there is enough data. The highly connected and semantically enhanced *FAIRMat* datasets are going to be an enabling resource for experimental, ML-based technologies.

1.4.6 Technologies and TRLs

FAIRMat chooses some of the most mature and most widely used existing services for mathematical data. Because these have so far been developed independently of each other, *FAIRMat* employs quite a number of different technologies. These are listed in in Figure 1.4.3, together with their technology readiness levels at the expected project start and end times, respectively. This list excludes those technologies that are well-understood and considered standard such as Linked Data standards like RDF or general purpose web service and database technologies like SQL. In the sequel, we briefly describe all technologies in detail.

#	Technology	Site	Initial TRL	Final TRL
	Languages and APIs used in WP2			
1	OMDoc	**FAU**	7	9
2	MMT	**FAU**	6	8
3	Mathematical schemas	**FAU**, **UPSud**	6	8
4	Modelica language	**CAE**, **EMS**	9	9
5	Emmo	**CAE**	6	9
	Services used in WP3			
6	MathWebSearch	**FAU**,**FIZ**	7	9
7	DiscreteZOO	**FAU**, **UL**	6	8
8	Recon	**CAE**	8	8
9	SageMath	**UPSud**	9	9
9(a)	Sage Explorer	**UPSud**	6	8
9(b)	DiscreteZOO SageMath package	**UL**	6	8
9(c)	MMT bridge	**FAU**	6	8
10(a)	Author Disambiguation	**FIZ**	8	8
10(b)	Citation Matching	**FIZ**	7	8
	Dataset hosting technologies relevant for WP4			
11	RADAR	**FIZ**	9	9
12	EuDML	**EMS**, **FIZ**	8	9
13	LMFDB	**CHA**	7	8
14	MathHub	**FAU**	6	8
15	swMATH	**FIZ**	9	9

Figure 1.4.3: Technology Readiness of the *FAIRMat* Technologies

WP2 develops our data representation standard. This will employ existing languages and implementations for them:

1. OMDoc [Koh06] is a representation format for symbolic data. It has been under (academic) development by the PIs of site **FAU** for over two decades.
2. MMT [MMTa] is a language definition and knowledge management framework based on OMDoc. It has been designed, implemented, and maintained by the PIs of **FAU** for 15 years.
3. Mathematical schemas are a standard for specifying the structure of mathematical datasets, together with an MMT-based implementation of that standard. They were developed by the PIs of **FAU** and **UPSud** in the OpenDreamKit project (2015–2019) and have been tested on the LMFDB, FindStat, and DiscreteZoo datasets and has been used with outside users.
4. Modelica is a modeling and implementation language for cyber-physical systems. It has developed since 2000 [MOb] and is widely used in academia and industry. The PIs of **CAE** and **EMS** are integral members of the Modelica community.
5. Emmo is an implementation of Modelica that provides advanced services for managing Modelica models. It is developed at **CAE**.

WP3 develops a set of services based for datasets using the representation standard from WP2. This will employ a variety of existing services, of which we list the most essential here:

6. MathWebSearch [PK11] is a search engine for full text and symbolic mathematical data. It was developed by the **FAU** PIs and has continuously been in active use in the zbMATH services for four years. It will be a part of **T3.4**.
7. The DiscreteZOO website framework generates the full stack (from database schema to web interface) for a mathematical record dataset from a schema theory. It has been tested on several datasets of combinatorial objects and is currently jointly developed at **FAU** and **UL**. It will be a central part of services for concrete data.
8. Recon [TH14] is an open-standard file format for time-series array data, e.g., as produced by Modelica-based simulation tools. It was developed by PIs at **CAE** and specifically designed for transmission across a network. It will be a part of **T3.6**.
9. SageMath (`sagemath.org`) is a Python-based open source system for computational (pure) mathematics, developed since 2005 by an international community of hundreds of researchers, teachers and engineers, with tens of thousands of users. The PI of **UPSud** is an integral member of the SageMath community. SageMath-based technologies will be used in various tasks including **T3.3**, **T3.5**, and **T3.7**. Multiple *FAIRMat* technologies have been developed by *FAIRMat* PIs in the form of SageMath packages. These are conceptually part of the SageMath ecosystem but are listed separately here in order to describe their role in *FAIRMat* and to ascribe them TRLs:
   (a) Sage-Explorer (`https://github.com/sagemath/sage-explorer`) is a prototype of live data browser for SageMath, developed by OpenDreamKit. It exploits the introspection features of the Python language and the semantic embedded into SageMath to offer rich interlinked views, letting the user query, explore, and navigate through the mathematical objects available in a SageMath session.
   (b) The MMT-Python bridge [MMTc] allows working with MMT documents from within Python and SageMath (including memory sharing). It was developed by the **FAU** PIs during the OpenDreamKit project.
   (c) The DiscreteZOO SageMath package enables working with databases of graphs through SageMath. It was developed at **UL**.
10. Within the zbMATH infrastructure developed and maintained at **FIZ**, several services have been developed and are in active use. These will be used in **T3.3** and **T3.4**. As in the previous item, they are listed separately even though they are conceptually part of the same technology:
    (a) Disambiguation techniques adapted to the specifics of mathematics have been in production use for a decade to disambiguate author information in zbMATH.
    (b) FIZ Karlsruhe uses and continuously refines state-of-the-art machine learning technologies for matching literature citations to mathematical research articles. Public APIs are used by a variety of users ranging from open-access digital libraries to commercial publishers.

WP4: Datasets hosts individual datasets on the *FAIRMat* platform. The datasets themselves are already listed in Figure 1.3.5. Here we only list existing dataset hosting technologies maintained by partner sites that will be reused in or linked with the *FAIRMat* platform.

11. The RADAR project has developed a sustainable generic infrastructure for research data management and archiving focusing on the long tail of science. In 2017, the project was successfully transformed into a production service.
12. The European Digital Mathematics Library (EuDML) is a collaborative digital library service that comprises distributed open publications from collections throughout Europe. It is maintained and developed by the EuDML Initiative, a not-for-profit association of eleven partners established under the leadership of the European Mathematical Society.
13. The LMFDB database has existed as a public service since 2008, and is used by mathematicians on a daily basis since then. In 2018, the underlying database has changed from MongoDB to PostgreSQL yielding much better transaction performance.
14. MathHub.info [MH] is a platform for active mathematical documents and flexiformal mathematical libraries (encoded symbolic data). The system uses GitLab for versioned storage, MMT [MMTa] for knowledge management, and TGView for graph visualization. The MathHub dataset contains $\approx$ 240 datasets ranging from formal theorem prover libraries over the MitM ontology, to semantic/active course materials and linguistic resources.
15. The mathematical software database swMATH is a free database of mathematical software with systematic linking of software packages with relevant mathematical publications. Various state-of-the art technologies

to extract automatically software information from publications are employed in its production since 2013 and have ever since lead to a continuous growth in content and features.

2 Impact

2.1 Expected Impacts

FAIRMat will massively improve the recognition and ability for Open Science and FAIR data sharing in the mathematical sciences and beyond. It will provide pivotal changes to the way how mathematical results are produced, maintained, communicated, and archived. It will contribute to a new generation of researchers that consider Open Science and interdisciplinary collaboration not a goal but a matter of course.

2.1.1 Expected Impacts Listed in the Work Programme

We describe how *FAIRMat* contributes to each of the three expected impacts listed in the work programme.

2.1.1.1 Expected Impact: Integrating research and service development As most mathematical services are written by the mathematical community, **mathematicians are already co-designing research and services in current practice**. The expected impact of the *FAIRMat* project is that we provide the infrastructure and in some cases funding to make that easier and make the results more accessible, findable, interoperable, and sustainable. By providing clients for widely used research systems (see **T3.5** and **T3.6**), **we make it substantially easier and cheaper for service developers to combine their implementation work with their traditional research**.

KPIs: To test this expectation, we will brief the participants of the "Summer of Math Data" development workshops (see **T5.2**) to keep track of their development investments and the outcomes achieved in terms of data and service features. We will evaluate these tallies and conduct interviews with the participants of the events about how much time they save by using our user-oriented services as a basis for their extensions. We expect that, even though the *FAIRMat* framework will still be in prototypical state at the time of these two "summers", the participants will already break even early on and will, during the project run time, see substantial savings of time and effort while improving their services.

2.1.1.2 Expected Impact: Supporting the objectives of Open Science **The *FAIRMat* project was specifically conceived to support Open Science in the mathematical sciences**, and its objectives directly feed into the objectives of Open Science. In particular, *FAIRMat* develops the foundations of FAIR Mathematics, implements a service infrastructure for Open Data in mathematical sciences, and evaluates it on a representative family of datasets and services. Thus, it enables enable scientific and industrial applications by **enabling users to openly access/search/reuse/interoperate with mathematical data from all scientific disciplines and sectors**.

For a concrete example of the value of Open Science in mathematics, consider the Fibonacci sequence (given by the recurrence equation $f(n) = f(n-1)+f(n-2)$): while it was originally of mathematical interest to describe the (idealized) reproduction of rabbits, it was soon discovered that this sequence governs the number of petals in certain kinds of flowers and is connected with the meters of traditional Indian poetry. It is one of about 325.000 sequences collected in the Online Encyclopedia of Integer sequences (OEIS) [OEIS; Slo03], and transdisciplinary interpretations of OEIS sequences abound but are non-obvious to find systematically. The *FAIRMat* project will integrate the OEIS and similar datasets into the EOSC and thus support the systematic cross-linking to other sciences.

KPIs: To evaluate the impact, we will track the number of databases on the *FAIRMat* server (both the ones described in WP4 and the ones contributed by mathematicians at the summer events) and the numbers of searches (finding) and downloads (access). We expect that the contributed databases matche the *FAIRMat* ones in number, and that the number of downloads goes into the dozens for each dataset from pure mathematics and higher for those from other disciplines.

2.1.1.3 Expected Impact: Opening up the EOSC ecosystem to new innovative actors The mathematical community is currently very under-represented on the EOSC: there is only one mathematical dataset and the EOSC is currently virtually unknown in the community. Through the outreach in the *FAIRMat* project (see WP5) and the integration of existing datasets (see WP4), we will **massively increase the number of mathematical datasets on and users of the EOSC ecosystem**.

We are very confident about the community of computational and applied mathematics embracing the EOSC as a useful resource — as they have eagerly embraced other open technologies like `http:arxiv.org`, LaTeX, or

SageMath in the past. In all cases, the mathematical community has directly contributed back – be it via arXiv submissions, LaTeX packages, and SageMath components. Moreover, mathematical users are highly innovative actors in these ecosystems as they are highly capable, motivated, and far-thinking while constrained by tight budgets that require them to automate as much as possible. In many cases, crucial technology innovations came from users in mathematics and related sciences such as LaTeX by Don Knuth and Leslie Lamport and the SageMath system by William Stein. **By directly working with these users and systematically integrating our services with their favorite systems, we allow this highly innovative community to adopt and contribute to the EOSC ecosystem.**

For a concrete innovation example, consider Enxhell Luzhnica's work [Luz16] (a bachelor's thesis later published as [LK16]), who extracted 50K generating functions of OEIS sequences and computed relations between these to obtain relations between the corresponding sequences. This work was only possible due to research contacts between the thesis advisor and the OEIS team. The *FAIRMat* project will make all kinds of datasets, including the OEIS, openly accessible, findable, and interoperable on the EOSC so that innovations like this can be done without personal contacts between researchers.

Dually, our experiences will provide feedback to existing systems about how to improve their FAIR readiness in the future. For example, in the Modelica ecosystem, deep findability will avoid the duplication of work and enhance model reuse and collaboration. This may require improvement in the Modelica language, and *FAIRMat* has the experts to credibly suggest such changes to the Modelica Association.

KPIs: We will measure the impact of opening the EOSC to innovative users by the number of external mathematicians uploading to or downloading from our server, and the new features in the *FAIRMat* services. Similarly, we will measure the number of datasets that are shared via our services and copied to EUDat (see **T3.8**).

We will measure innovation qualitatively by listing innovative solutions enabled by *FAIRMat*. An example of such innovation is a package manager and cross-model-set search for the Modelica ecosystem, see **T3.6**, which have been consistently reported missing by the Modelica user community. Another is the integration of SageMath with live mathematical datasets and the use of datasets for persistent memoization (i.e., the creation and use of ephemeral datasets during computation), see **T3.5**.

2.1.2 Exploitation of results, enhancing innovation and knowledge production capacity

FAIRMat will act as a catalyst for innovation and knowledge production. Generally speaking, it will improve the capacities of multiple groups of ESOC ecosystem stakeholders:

1. *Pure and Applied Mathematicians* can directly work with the datasets and services provided by the *FAIRMat* project. Moreover, the *FAIRMat* platform gives dataset providers an easy framework for making their results available and academically recognized. This will also free scarce resources of highly trained researchers that are currently tied up by the overhead of data sharing.
2. *Natural and Social Scientists* can do the same for their interests, e.g., searching possibly connected phenomena via the count sequences (e.g. petal numbers in flowers related to rabbit populations and crystal faces).
3. *Educators* will have an easy-to-use resource for mathematical examples (the *FAIRMat* datasets) that are integrated into mathematical education infrastructures like CoCalc [CC] (via the SageMath interfaces from **T3.5**: CoCalc is based on SageMath).
4. *Funding agencies and Hiring Committees* have a central framework to judge direct contributions and reuse patterns in mathematical data. This will ultimately better integrate the recognition of data contributions into the academic reputation economy and make data production a more attractive proposition for junior researchers.
5. *Industry* — in particular engineering companies — can use mathematical model databases and related services. Industrial stakeholders will be directly involved in the development of the data standards and framework, so that the services will be exactly tailored to their specific needs as well as to the needs of the scientific community. Moreover, this will allow short time-to-market and will facilitate the technology uptake (see **T3.5**)

More concretely, in the next table we describe different market needs and how we can leverage *FAIRMat* results to address them:

Market need	How *FAIRMat* will address it
Coverage gain	The *FAIRMat* framework will give users access to a uniform, interoperable collection of mathematical data and services whose scope extends over all data.
Infrastructure for sharing and archiving	Mathematical datasets are often lost because there is no infrastructure to host and archive them properly (and university resources go away). *FAIRMat* directly addresses this need.
Low scaling costs	The standardized and open architecture of *FAIRMat* brings affordability: individual researchers, small communities, and small organizations can gain access to equipment otherwise only affordable by the largest companies.
Human Resources	By strengthening open source infrastructures, *FAIRMat* will enhance the potential of research projects to attract the best minds from a wide pool of people to solve a problem.
New applications and features	*FAIRMat* will generally improve existing tools and their FAIR readiness and push these improvements back into the systems' development pipeline.
Short time-to-market (TTM)	*FAIRMat* will speed up development of tools by allowing stakeholders to communicate and share data more efficiently.
Ease of use	First-time experiences are crucial to gain acceptance. *FAIRMat* will design an ergonomic multi-user web-based graphical user interface, following web standards to best support a large array of browsers, including cell phones and tablets. We will explore opportunities for integration in interactive boards, as an aid for teaching and collaborative research.

This analysis directly shows the road towards exploitation: in our dataset work package WP4 and our community outreach work package WP5, we budget extensive resources to increase the visibility and use of both the general EOSC infrastructure in general and the *FAIRMat* service in particular. **By EOSC-publishing a collection of important datasets ourselves, deeply involving community multipliers in our research, and integrating our services with existing widely used systems, we jump-start the realization of the impacts predicted above.**

KPIs: Being inherently a catalyst, the exploitation is hard to measure. We will reuse the KPIs from Section 2.1 to measure the adoption of EOSC and our *FAIRMat* services and to evaluate the results of our exploitation efforts.

2.1.3 Integration into and Impact on the EOSC

***FAIRMat* is systematically designed to deliver a succinct service that can be readily integrated with the EOSC Hub.** As described in WP3, our entire software infrastructure will be designed to enable the smooth porting to official EOSC servers. That includes in particular the compatibility of *FAIRMat*'s deep FAIR services with the existing shallow FAIR services on the EOSC Hub, such as B2Handle, B2Share, etc.

Deep FAIR services are a critical prerequisite for large-scale adoption of the EOSC, especially for transdisciplinary reuse because they allow much more meaningful interaction between researchers. Especially for large and complex datasets — as typical for, but not limited to the mathematical sciences — it is important that researchers can find/access/reuse/interoperate with dataset fragments and moreover do so using automated tools. This can only be accomplished if the semantics of datasets is described in machine-accessible ways, i.e., by using technologies as envisioned by *FAIRMat*. For example, with shallow FAIR services only, a researcher has virtually no chance of even finding a relevant dataset because the shared version of the dataset uses encodings that make them non-understandable to a search engine. Only by attaching machine-readable semantics, i.e., by enabling deep FAIR services, can a search engine determine which datasets could be interesting for the potential reuser. This argument is all the more important, the more scientifically remote from each other the producer and the reuser of a dataset are. Already within pure mathematics, it is difficult for experts to reuse datasets from other areas with machine support for understanding the details of those datastes. For transdisciplinary reuse, the situation is correspondingly harder.

KPIs: The evaluation criterion for this impact is straightforward: it is the possibility and easiness to deploy the *FAIRMat* services on the EOSC Hub, which will be documented in the respective deliverables of WP3.

2.1.4 Obstacles and Framework Conditions

The following barriers to impact will be addressed and overcome using the mitigation strategies provided. These are distinct from the risks to project delivery detailed in Section 3.2.9.

Barrier description	**Risk level**
Adoption of EOSC Hub	Low
If the EOSC Hub were to be ineffective in achieving adoption of the scientific community, it would not be able to host the *FAIRMat* data and services.	
Contingency Plan: The *FAIRMat* project will help EOSC Hub adoption by integrating high-quality data and services actively advocating its use. Should the EOSC Hub fail to attract significant users anyway, the *FAIRMat* services and data can be hosted on existing community platforms and thus achieve the intended impact regardless.	
Adoption of EOSC-based Mathematical Services	High
A major concern in any proposal of this kind is that the resulting tools will not be adopted by users. This is a particular concern with a 'tradition-based' community such as mathematicians.	
Contingency Plan: This project employs multiple strategies to avoid this: (1) *FAIRMat* is based on prior work, which *already* has users. (2) *FAIRMat* will be integrated into the EOSC Hub to enhance its appeal. (3) We will form an end-user group (see Section 3.2.6) at the beginning of the project that will include representatives from different disciplines and sectors and will provide valuable advice on real user needs throughout the project. (4) We will work towards good integration of our services with existing non-EOSC technologies, especially if that meets user needs (e.g., by being more flexible or more lightweight). (5) Mathematical datasets are often living objects, and we will support continued synchronization with updated datasets to retain users.	
Dominance of existing Data Services	Medium/High
Much mathematical data is already managed in — insular, but well-established — platforms, which also provide substantial services. It is hard for any alternative, both for the EOSC Hub in general and for *FAIRMat* in particular, to compete with these.	
Contingency plan: First of all, *FAIRMat* significantly increases the competitiveness of EOSC Hub compared to established platforms. Additionally, we will engage with users and attract new users right from the beginning of the project in order to understand their requirements and design our services accordingly. We will create an international advisory board (see Section 3.2.5) with senior members whose opinion will affect the adoption of our platform. This will allow us to coordinate with the related research activities within and outside of Europe and to promote our framework internationally.	
Brexit	Medium/Low
CAE is in the UK and would be affected by a no-deal Brexit.	
Contingency plan: CAE can rely on the underwrite guarantee created by the UK Government: "UK participants that receive Horizon 2020 funding from the European Commission or have submitted a bid before EU exit and are notified of their success after exit will be covered by the underwrite guarantee, for the lifetime of the projects" See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/766510/horizon-2020-government-overview-december-2018-update.pdf.	

Table 2.1.2: Barriers to Impact

Sustainability and Follow-Up Financing In the long run, we expect our services to be hosted by the EOSC, akin to the current hosting of the B2 services. Our services will be designed in a way that only minimal efforts are needed to integrate them into the EOSC. Therefore, only little follow-up funding will be needed to perform the integration. During the transition period of 1-2 years, the services can continue to be hosted by the project partners using the prototype server built during the project.

2.2 Measures to Maximize Impact

The overall objectives of the dissemination and exploitation strategy are based on the project's core values, which are to improve the productivity of researchers in mathematics and connected fields by providing them with the easy and uniform access to mathematical datasets and novel and innovative services for data.

Dissemination goal	Target audience	Dissemination method	Timeframe and frequency
Project identity and profile	T1-T6	Website; flyers/leaflets; videos.	Throughout project, continuous
Broad dissemination	T3, T4, T6, T7	News feed on the Web Site; press releases; information database; social networks and platforms;	Throughout project, quarterly
Knowledge transfer, information exchange	T1,T2, T3	Organization of outreach workshops at math community meetings like the annual DMV meeting; organizing two "Summer of Data" events which bring together users, framework implementors and data providers; technical workshops at framework conferences like the Conference on Intelligent Computer Mathematics; see **T5.2** for details.	Throughout project, continuous
Uptake of the EOSC Services by new users	T2-T4	Integration of datasets (see WP4); white papers; scientific trainings for other scientific communities/projects (see WP5); presentations at international conferences; at least two new data carpentry lessons.	Mo18-, at relevant milestones
Sustainable development beyond the project	T1-T4	Policy events; white papers; participation at conferences.	Mo18-, at relevant milestones

Key for the "Target Audience" column:

- T1 Scientific community in mathematics and related fields (experienced researchers, under-/graduate/post-graduate students)
- T2 Scientific community in other disciplines
- T3 Other relevant European and national initiatives and projects
- T4 Industrial end-users
- T5 Standardization agencies
- T6 Civil society
- T7 Public at large

Table 2.2.2: Dissemination and exploitation plan

2.2.1 Dissemination and exploitation of results

2.2.1.1 Dissemination during the project The concrete dissemination and exploitation strategy will be presented in the dissemination and exploitation plan, prepared by the Coordinator within the specifically designed WP1 and implemented with the help of all partners. The planned activities will bear in mind the project's scientific and societal impacts, and build throughout the project to ensure that stakeholder communities (1) are fully aware of the project and its potential benefits, (2) engaged in integration of their datasets and collections into the EOSC in their professional activities, and (3) contribute to the sustainability and improvement of mathematical data on the EOSC.

We summarize the dissemination activities and how they will help to achieve the expected impact among our stakeholders and target audiences in Table 2.2.2.

Long term sustainability and exploitation The natural interest of the consortium is to ensure the permanent sustainability of the *FAIRMat* datasets and services after the completion of the project. The **maximization of long-term impact is baked into the proposal in several ways**:

- We will submit out data representation standard for ISO standardization, see **T1.3**.
- We will make the EOSC Hub-ready deployment of the *FAIRMat* services the crowning task of WP3.
- WP4 was designed to provide an impact-maximizing selection of datasets that we will make available via the *FAIRMat* and EOSC-Hub infrastructure.

- WP5 is designed to initiate and deepen research collaborations that will endure beyond the project duration.

Moreover, **the partners are committed to post-project efforts** including the following activities:

1. We will continue the dissemination to scientific community and industrial stakeholders through participation to international conferences and publications, including software demonstrations during the conferences.
2. We will train students and colleagues in using the *FAIRMat* technologies to publish and work with mathematical datasets.
3. We will expand the *FAIRMat* user base by continuing the research collaborations with existing users and identifying new scientific (specifically from neighbouring fields) and industrial users.
4. We will apply for funding at European and national levels for related projects, in particular to deepen the integration with the EOSC and with national research data initiatives. A particularly promising venue is the recent German NFDI initiative; the PIs of **FAU** and **EMS** are part of the "Math4NFDI" consortium that is in the process of founding itself.

The details of these will be worked out by the *FAIRMat* consortium in task **T1.4**: Ensuring Sustainability of EOSC Math Data/Services and reported on in the Sustainability plan (see D1.4). By the end of the project, we expect that the framework will have been well-developed and tested so that financial support will become less important and the curation of the datasets can be organized by the community and its professional organizations — having the **EMS** as part of the consortium will facilitate this.

Exploitation The aims of the *FAIRMat* project are to make mathematical data FAIR and useful to fellow mathematicians, scientists from other disciplines, European industry, and the general public. Our main **exploitation** regime is to ensure that this will happen with maximal effectiveness.

Therefore, software developed in the *FAIRMat* project will be open source, and all data will be licensed under open licenses where possible.

Open access policy and data protection. The consortium will comply with the Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. We will generate a detailed description of the data sources with specifics pertaining to data management (metadata standards, policies for access and sharing and for reuse and distribution, plans for archival and preservation, with accompanying deadlines). This information will be presented in the Data Management Plan, to be delivered within the first six months after the project start and subsequently updated throughout.

All scientific publications produced in *FAIRMat* will be either published in open access journals or self-archived using research data repositories. In addition, we will make all experimental data needed to reproduce/validate the results from scientific publications available through our own data sharing infrastructure.

Intellectual Property Rights Management. IPR management will be described in detail in the Consortium Agreement (CA), which will describe all issues regarding the IPR, confidentiality, know-how, rights on exploitation, the rights and obligations of the each partner. The CA will be prepared by the Coordinator, and then signed by all partners before the start of the project.

Access rights to foreground and background needed for the execution of the project shall be deemed granted, on a royalty-free basis, as of the date of the grant agreement entering into force. Methodology, documents, know-how, software, and tools will be available to all in order to achieve the project objectives during the project lifetime.

Most of the project results will have joint ownership due to a highly collaborative nature of the project. The CA will specify the terms of the resulting joint ownership, i.e., assignment of shares between joint owners, conditions of use, exploitation and management of jointly used IP.

The CA will also outline rules for publication procedures to ensure that IP can be protected while minimizing publication delay.

The costs related to IPR (including those related to protecting results) and dissemination (i.e., 'gold' open access publications) are included in the project budget of each participating organization.

2.2.2 Communication activities

Our intention is to increase the attractiveness of mathematics among young generation and females in particular as well as to improve the impact and maximize the visibility of the project activities for mathematical data and services on the EOSC. The following strategic access points will be used to maximize visibility:

1. an online presence that explains the *FAIRMat* concept and its applicability in layman's terms and offers significant information (website, social networks, Youtube, press releases),
2. collaboration with other relevant European and national projects (existing and new ones),

3. collaboration with national mathematical societies; the European Mathematical Society is already a partner in the *FAIRMat* consortium),
4. presentations/demonstrations at partner institution-specific, locally organized 'science holiday' and 'days of science',
5. involvement in workshops/conferences on e-infrastructures and broad mathematical topics.

3 Implementation

3.1 Work Plan — Work packages, deliverables

3.1.1 Overall Structure and Timing

The **overall structure of the work plan** is very simple and follows immediately from the methodology described in Section 1.3.4:

- WP1: Management covers all management and administrative tasks.
- WP2: Framework builds the data representation framework. This includes the standardization of the data formats, the specification of the core APIs (which serve as the foundation for building the services), and legal aspects.
- WP3: Services builds the services. This includes both the design and implementation of the software, as well as the setup of appropriate hardware and the deployment of the services.
- WP4: Datasets seeds our infrastructure by deploying a collection of datasets on it. This includes at least the datasets specified in the respective deliverables but will likely involve more datasets based on feedback from the community.
- WP5: Dissemination and Community Organization handles all community outreach and advertisement within new user communities.

The **list of work packages** is given in Figure 3.1.1.

WP	Title	FAU	UPSud	CHA	UL	CAE	FIZ	EMS	total
WP1	Management	*18*	1	1	1	1	1	10	**33**
WP2	Framework	*22*	5	5	5	5	11	12	**65**
WP3	Services	28	16	15	*27*	17	3	3	**109**
WP4	Datasets	12	11	12	*13*	10	12	3	**73**
WP5	Dissemination and Community Organization	10	3	3	8	3	9	*8*	**44**
totals		**90**	**36**	**36**	**54**	**36**	**36**	**36**	**324**

Efforts in PM; WP lead efforts light gray italicised

Table 3.1.1: List of Work Packages

The above linear listing of work packages WP2-WP5 is not meant to imply that they are processed in order. In fact, all work packages will proceed in parallel and be active throughout the project duration.

Each work package consists of multiple tasks, each of which results in one deliverable. The resulting list of deliverables is given in Section 3.1.3.

A Gantt chart showing the **timing of the work packages and their components** is given in Figure 3.1.2 on page 50. A Pert chart showing the **interrelation of the tasks** is given in Figure 3.1.1 on page 33.

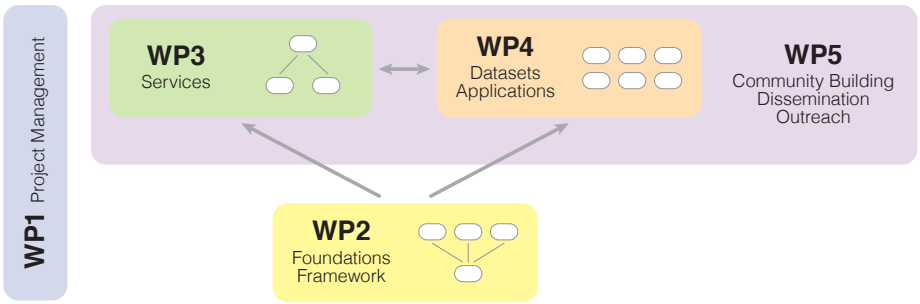

Figure 3.1.1: Pert Chart (Dependency Structure) of the *FAIRMat* Project

3.1.2 Detailed Work Description

Work Package 1: Project Management								**Start:** 0
Site	***FAU***	**UPSud**	**CHA**	**UL**	**CAE**	**FIZ**	**EMS**	**all**
Effort	*18*	1	1	1	1	1	10	**33**

Objectives

This work package will establish and maintain an effective contract, project, and operational management approach, ensuring (i) an effective and timely implementation of the project, (ii) quality control of the results, (iii) risk and innovation management of the project as a whole as well as (iv) timely and necessary interaction with the EC and other interested parties.

Description

The project will be managed by **FAU**, which has extensive experience in administrating and leading EU funded and national projects. The coordinator, together with the WP leaders, will be responsible for monitoring WP status, coordination of work plan updates and annual internal progress reports. The project management structure and roles of partners in the consortium are presented in Section 3.2.

T1 Project and financial management; Sites: **FAU** (lead), **CAE**, **CHA**, **EMS**, **FIZ**, **UL**, **UPSud**
The task includes the following activities

- Preparation and Distribution of the Consortium Agreement;
- Setting up the project website, intranet and communication procedures for effective communication;
- Organisation of project review and progress meetings;
- Establishment and maintenance of external contacts (with the EC, other relevant national / EU projects, other academic and industrial stakeholders) to organise transfer of knowledge, present and promote project results;
- Progress and Financial Reporting to the EC;
- Data and IPR Management will be managed in accordance with agreed rules stated in the Consortium Agreement and in accordance with the Data Management Plan (D1.2).

This task will be led by the coordinating site with minor contributions from all partners.

T2 Quality assurance and risk management; Sites: **FAU** (lead), **CAE**, **CHA**, **EMS**, **FIZ**, **UL**, **UPSud**
A quality assurance plan will be established to ensure coherent and sufficient quality of the work and results. The plan will be developed by **FAU**, involving all partners, and will be followed up regularly. In addition, the CT with support from the SC will establish and review annually a risk management plan and self-assessment to ensure that technical barriers / potential risks are identified and corrective measures are put into place on time (D1.3).
This task will be led by the coordinating site with minor contributions from all partners.

T3 Preparing ISO Standardization; Sites: **EMS** (lead), **CAE**, **CHA**, **FAU**, **FIZ**, **UL**, **UPSud**
It is a central aim of the *FAIRMat* project to set open standards for all three kinds of mathematical data. To maximize the impact of the *FAIRMat* standard developed in **T2.4**, we will develop it into an ISO standard. The committee ISO/IEC JTC 1 “Information technology”, which has already standardized base technologies like MathML, and/or the ISO/TC 37 “Language and terminology”, which has standardized ontology languages like OntoIOp, are likely venues for this.
In this task, we carry out the administrative and political work as well as the work of all necessary revisions necessary to submit our standard. We aim to file the submission at the end of the project duration.
This submission process will be led by **EMS** with minor contributions from all sites.

T4 Ensuring Sustainability of EOSC Math Data/Services; Sites: **EMS** (lead), **CAE**, **CHA**, **FAU**, **FIZ**, **UL**, **UPSud**
The aim of the *FAIRMat* project is to kick-start the development of mathematical data and services on the EOSC by standardizing a framework and providing initial datasets and services. After the project it will be easy (technically tractable, well-documented, and well-advertised) to extend the EOSC datasets and services. In fact it will be simpler – e.g. for Ph.D. students in Mathematics who create mathematical datasets or services – to integrate them into the *FAIRMat* framework at EOSC than building/hosting their own. In essence

the proposed *FAIRMat* framework will provide a sustainability solution for individual Math researchers, but this long term sustainability needs to be organized and disseminated to be successful.
The aim of this work package is twofold: to plan for making the *FAIRMat* framework and infrastructure itself sustainable – i.e. to organize stewardship for the time after the project, and to plan a sustainability framework for individual mathematical research contributions. The latter part includes *i*) assessing the technical tractability of importing datasets into the *FAIRMat* data framework and integrating services into the EOSC. *ii*) organizing (and testing) suitable documentation and tutorials, and *iii*) organizing the respective outreach activities. Tasks **T1.3** and **T2.4** feed into this activity, but are distinct.
This task will be led by **EMS** with minor contributions from each partner.

Deliverables:
D1.1 (Due: 2, Type: DEC, Dissem.: PU, Lead: FAU) *Basic project infrastructure (websites, wikis, issue trackers, mailing lists, repositories)* ⇝**M1**
D1.2 (Due: 6, Type: R, Dissem.: PU, Lead: FAU) *Data Management Plan* ⇝**M1**
D1.3 (Due: 15, Type: R, Dissem.: CO, Lead: FAU) *Internal Progress Report year 1, including risk management and quality assurance plan* ⇝**M2**
D1.4 (Due: 24, Type: R, Dissem.: CO, Lead: EMS) *Sustainability Plan* ⇝**M2**
D1.5 (Due: 24, Type: R, Dissem.: CO, Lead: FAU) *Internal Progress Report year 2* ⇝**M3**
D1.6 (Due: 36, Type: R, Dissem.: CO, Lead: FAU) *Final Achievement Report* ⇝**M4**

Work Package 2: Framework								**Start:** 1
Site	***FAU***	**UPSud**	**CHA**	**UL**	**CAE**	**FIZ**	**EMS**	**all**
Effort	*22*	5	5	5	5	11	12	**65**

Objectives

The objective of this work package is to lay the technical, organizational and legal foundations for FAIR mathematical data on the EOSC. Technically, this involves the design and standardization of the data representation framework underlying all services. This must satisfy the following requirements:

R1 It must provide a coherent integration of symbolic, concrete, and linked data in a way that makes data from different sources or of different kinds interoperable.

R2 It must define APIs that allow the realization of semantics-aware FAIR services while making it easy for data providers to make their data available.

R3 It must allow for the optimized representation of large-scale datasets. In particular, the formating standard may not get in the way of representing objects in the most scalable way (which may require, e.g., binary encodings or compression).

R4 It must define appropriate legal policies for licensing, sharing, and governing.

R5 It must allow for multi-level-grained provenance information that enables reuse in a way that satisfies mathematical rigor.

R6 It must support datasets that grow over time. Contrary to other disciplines, mathematical datasets are often only finite subsets of infinite sets and are continually updated, often collaboratively, as new elements are found.

Description

This work package culminates in **T2.4**, which delivers the data representation standard. The other tasks feed into it by developing certain components of the standard that require separate attention:

- **T2.1** develops the standardization needed to make data semantically accessible.
- **T2.2** develops the standardization of all metadata, including provenance, versioning, and licenses.
- **T2.3** develops all legal aspects.

These three tasks are formally delivered at Month 12, and the overall standard at Month 24. The work package is led by the coordinating site **FAU**, specifically by Florian Rabe, who has worked specifically on formal knowledge representation frameworks and their implementation over the last 15 years.

T1 Reference Library of Data Types & Codecs; Sites: **FAU** (lead), **CAE**, **CHA**, **UL**, **UPSud**

Specifying the semantics of mathematical data is hard. For symbolic data, this has been solved by the use of meta-logical frameworks and has been implemented at scale in the MMT system at **FAU**. For linked data and concrete, the use of ontology languages like OWL resp. database schemas is common; but these only offer general purpose datatypes like numbers, strings, and untyped lists, which is too weak for the complex datatypes that pervade mathematical sciences like polynomials, multidimensional arrays, graphs, towers of algebraic structures (e.g. matrices over polynomials over algebraic extensions over finite fields), physical quantities, or numbers with error intervals. These have to be *encoded* in terms of the low-level datatypes. If these encodings are not described in detail, the data is not reusable. For example, Kohonen's lattice dataset uses 5 encoding steps: lattices are encoded as graphs with canonically labelled nodes, the graphs as adjacency matrices, the adjacency matrices as bit vectors and the bit vectors as `digraph6` strings (similar to `base64`), and finally the entire file containing many lattices is gzipped. Similar steps are needed for the graph datasets. Even when these encodings are documented, they are tedious and error-prone, and make difficult any automated processing needed for data validation, reproduction, or machine learning.

In the OpenDreamKit project, the FAU group has developed a systematic solution by annotating datasets with formal schemas that specify both the high-level mathematical type and the encoding function.

In this task, we expand on these efforts. We standardize a fixed set of mathematical datatypes (such as the one mentioned above) that subsumes at least all datatypes occurring in the datasets of WP4. Moreover, we standardize encodings for these datatypes, again subsuming those used by practitioners in building their datasets. The biggest subtask here is in surveying the practically used datasets and making sure our standard is comprehensive enough. For example, just in the LMFDB database, we already encountered three different encodings of the type of large integers (i.e., integers too large to fit into numeric integer types).

Finally, we develop reference implementations for these codecs in various programming languages. This allows datasets annotated by our formalism to be type-checked and decoded automatically. The latter is critical for allowing other researchers to import and reuse a dataset or to build crawlers that read all datasets and build an index of mathematical objects. The services developed in WP3 will heavily depend on these codecs.
The task is led by **FAU**, which already led the initial efforts in the OpenDreamKit project. **UPSud**, **CHA**, **UL**, and **CAE** will provide feedback from a user perspective, in particular ensuring the comprehensive coverage.

T2 Ontology for Metamathematical Data Aspects; Sites: **FIZ** (lead), **FAU**, **UL**

Mathematical data are of an inherently diverse nature. We standardize a metadata ontology. By "metadata" we mean any additional data that talks about the data itself. This goes beyond attaching standard metadata like authors, dates, institutions, and descriptions. It would require additional semantic information which may ideally cover, e.g., the various types of mathematical software and the intrinsic relations to their output. Another layer of metamathematical information is the interface to other exact sciences, which employ mathematical models.
Importantly, it will include provenance data, which is critical for mathematicians to communicate the mathematical trustworthiness of data.
Notably, in mathematics, these metadata may often be mathematical data itself. This is most apparent for provenance data. An example of this is symbolic data, where provenance can span multiple layers. We have the case where the provenance for a single datum (e.g. in the form of a theorem or lemma) is given by a proof and its metadata, up to whole libraries, gigabytes in size, that are generated from earlier versions by knowledge compilation.
These enhanced metadata will allow to employ functions of like MathHub by **FAU** or RADAR by **FIZ** and facilitate the development of the services in **T3.3**.
This task will be led by **FIZ**, which has extensive experience due to their development and maintenance of the zbMATH service. **FAU** will provide feedback regarding usability and integration into the overall standard.

T3 Legal Aspects & Licensing; Sites: **EMS** (lead), **CAE**, **CHA**, **FAU**, **FIZ**, **UL**, **UPSud**

Generally speaking, there is little resistance in the mathematical community against the principles of Open Science; in fact, many mathematicians have a strong opinion in favor of it. In addition, ethical or intellectual property issues rarely arise around datasets. Nevertheless, the issue of licensing them tends to remain under-appreciated: most authors do not want to be bothered by licensing questions or are under the impression that datasets are accessible even if unlicensed.
In this task we tackle all legal issues that go beyond what is provided by the EOSC already, i.e., in particular any legal issues specific to mathematics. This includes the development of a data sharing policy and license recommendations.
This task will be led by **EMS**, which has the stakeholder relations and institutional support to develop wide-coverage legal policies and guidelines. All other sites will provide feedback from the stakeholders' perspectives.

T4 Standardization of Data Format and APIs; Sites: **FAU** (lead), **CAE**, **CHA**, **EMS**, **FIZ**, **UL**, **UPSud**

In this task, we develop the overall standard for representing mathematical data. The critical requirement is to allow datasets for any kind of data with accessible semantics. Our standard will subsume existing standards for specific kinds of data and systematically connect them to obtain a coherent whole. Incidentally, this allows reusing the multiple competing standards that already exist and that users do not switch away from easily.
For symbolic data, this is based on the OMDoc representation language and its implementation MMT, both built at **FAU**. In particular, we use the concept of meta-logics to ascribe the semantic to symbolic datasets. For example, for Modelica models, this involves representing Modelica itself as a meta-logic in MMT and then reusing it when representing individual models.
For linked data, we use existing standard formats based on, e.g., OWL or RDF. The primitive objects in these format are usually assumed as atomic identifiers (given URIs), which are treated as black boxes. Symbolic data has to be used to formally describe the internal structure and semantics of these objects. To couple the linked data to its symbolic metadata, we use the same URIs in both the linked and the symbolic data. Abstractly, these can be understood as meta-logics represented in MMT, combined with concrete encodings for serialization.

For concrete data, we use record and array representations based on standard formats such as JSON or SQL. As for linked data, we are flexible with respect to the specific format because we use codecs anyway to connect the data to its semantics given as symbolic data. For example, we can easily subsume the standard developed by **CAE** for representing simulation results. Critically, we make use of the codec framework developed in **T2.1** in order to systematically link the concrete datasets to their mathematical representation. All sites jointly develop the standard with work package leader **FAU** coordinating. **EMS** also contributes heavily as it will coordinate the subsequent ISO standardization in **T1.3**.

Deliverables:

D2.1 (Due: 12, Type: R, Dissem.: PU, Lead: FAU) *Datatypes and Codecs* ⇝**M2**
The specification of datatypes and codecs from **T2.1** with tutorials, prepared as a document.

D2.2 (Due: 12, Type: DEM, Dissem.: PU, Lead: FIZ) *Metadata Ontology* ⇝**M2**
The ontology for all meta-mathematical data from **T2.2**, prepared as a document.

D2.3 (Due: 18, Type: DEM, Dissem.: PU, Lead: EMS) *Legal Issues for Math Data/Services in EOSC* ⇝**M2**
All legal policies and guidelines from **T2.3**, prepared as a document.

D2.4 (Due: 30, Type: DEM, Dissem.: PU, Lead: FAU) *Standard Math Data on EOSC* ⇝**M3**
The overall standard from **T2.4**, prepared as a document.

Work Package 3: Services								**Start:** 1
Site	**FAU**	**UPSud**	**CHA**	***UL***	**CAE**	**FIZ**	**EMS**	**all**
Effort	28	16	15	*27*	17	3	3	**109**

Objectives

The objective of this work package is the design, implementation, and deployment of the semantics-aware services for mathematics based on the deepFAIR framework mathematical data developed in WP2. These must satisfy the following requirements:

R1 They must provide significant added value compared to the state of the art in ad hoc disparate services that are currently around.

R2 They must be user-oriented and be designed to maximize community adoption and innovation potential.

R3 They must scale to the order of typical mathematical datasets, ranging from MB to TB.

R4 They must not be arbitrarily limited to data from mathematics and also support mathematically structured data from other disciplines.

R5 They must anticipate and prepare the integration with existing EOSC services, particularly, be compatible with the B2ACCESS (secure authentication and authorization), B2HANDLE (assigns PIDs), B2SHARE (store and share small-scale research data), and B2FIND (discovery service based on metadata steadily harvested from research data collections) services, and possibly with B2DROP (secure and trusted data exchange), B2NOTE (data annotation), and B2STAGE (transfer of data to computation resources).

Description

The tasks in this work package can be split into groups as follows:

- Basic tasks provide the technical foundation: **T3.1** sets up the hardware and **T3.2** provides the core software on which other services built.
- Tasks **T3.3** and **T3.4** enable the FAIR principles of Accessibility and Findability, respectively. (Reuse is already guaranteed by the framework developed in WP2, which makes the licenses and semantics accessible.)
- We devote three tasks to providing advanced user-oriented services. These respond to the needs of specific user communities by making our software infrastructure more interoperable with existing systems. Concretely, **T3.5** builds a *FAIRMat* client inside the widely used computation system SageMath. Similarly, **T3.6** provides a client inside OpenModelica. In both cases, we demonstrate how to offer user services directly from within their natural environment, e.g., by allowing a search of the EOSC datasets from within their client. **T3.7** provides support for mathematical database views that allow reusing specific fragments of a dataset when (as is often the case) the dataset is too big to download as a whole.
- Finally, **T3.8** is a cross-cutting task than ensures EOSC compatibility with the goal of eventually deploying our services on the EOSC.

Except for all tasks depending on the basic ones, all tasks are mostly independent of each other. The work package lead **UL** only has to provide minor coordination of APIs, with most of the lead work relegated to the respective task leader.

Each task has one deliverable, which consists of the running hardware or software, the open source code, and a service description document that includes user-oriented parts like tutorials and usage examples.

T1 **Basic Infrastructure**; Sites: **FAU** (lead), **UL**

In this task, we purchase a server and set up the basic infrastructure on it. Here by “infrastructure”, we mostly mean *software* infrastructure — the hardware will be off the shelf.

We immediately provide shallow FAIR services, i.e., the upload/download of datasets, assignment of identifiers/DOIs, metadata annotation, license annotation, and versioning. Many of these services are already provided by EOSC, and we take care to ensure compatibility, with the goal of eventually replacing our shallow FAIR services with the official ones at the EOSC.

However, we will also explore to whether sightly different service designs might fit the needs of users, specifically in mathematical sciences. We do not believe that GitHub by itself is an appropriate model for the EOSC, but we believe that some advantages can be combined, most importantly the ease of use. For example, we could develop a model where users can opt-in to have their dataset automatically copied to GitHub.

This task will be led jointly by **FAU**, which will host the hardware, and **UL**, which will lead the service development.

T2 Reference Implementation and Semantic Validation; Sites: **FAU** (lead), **CAE**, **UL**

In this task, we develop a reference implementation of the standard developed in WP2. This will take the form of an API that can be easily reused in other applications, most importantly in the other services. This ensures that all services can work with the same data model.

This API will reuse existing APIs, most importantly the MMT API for symbolic data and encodings developed at **FAU**. Being based on MMT, this API will be written in Scala and offers a multitude of interfaces including HTTP, Java, and Python. This allows in particular the integration with, e.g., OWL API for linked data and the Emmo system for Modelica developed at **CAE**.

The most basic feature provided by this reference implementation is format validation. This allows, e.g., to type-check all encoded objects against their semantic mathematical type, thus avoiding encoding as described in Section 1.3.1.

This task will be led jointly by **FAU**, which leads the standard development and maintains the MMT API, and by **UL**, which will lead the implementation of service-specific interfaces on top of the basic API. **CAE** will provide support for subsuming Modelica.

T3 Human Accessibility and User Interface; Sites: **UL** (lead), **CHA**, **FAU**, **FIZ**, **UPSud**

In this task, we set up a portal to our shallow FAIR services. This will be a from-scratch implementation but building on previous experience on similar user systems such as MathHub by **FAU**, RADAR by **FIZ**, the LMFDB interface in which **CHA** was involved, and the Sage interfaces developed by **UPSud**.

In addition to a human-oriented user interface for the shallow services, we will develop two advanced accessibility services that are specific to the needs of users in the mathematical sciences and engineering.

Firstly, we develop accessibility services for users with disabilities, e.g., to read out mathematical datasets for blind users. This critically requires the codec framework developed in **T2.1** as reading the encoded data is useless to humans — only the decoded data yields a mathematical object that can be communicated to a user.

Secondly, we adapt the existing visualization components developed by the partners to make mathematical data accessible in ways more enticing and practical for human users. This includes the browsing and management of large symbolic datasets (MathHub, **FAU**), the visualization of large graphs of mathematical objects (TGView(3D), **FAU**), the semantic interaction with symbolic data (MMT, **FAU**), property-based presentation of mathematical objects (Sage-explorer, **UPSud**), native visualization of mathematical objects within computational systems (Sage, **UPSud**), and the exploration of datasets of mathematical objects via their mathematical invariants (DiscreteZOO, **UL**; LMFDB, **CHA**).

This task will be led by **UL**. Other sites contribute as indicated above.

T4 Indexing, Search, Finding Connections; Sites: **FAU** (lead), **CHA**, **FIZ**

We develop a service that crawls all datasets on our software infrastructure and produces a semantic index. This includes a substitution tree index for all symbolic data, a value index of all basic objects occurring in concrete (e.g., an index of all integers with their factorizations (if known), or an index of all polynomials, or graphs), and a graph index (triple-store) of all linked data (including the metadata of all datasets).

We use this index to build an efficient search service and integrate it into the user interface. This will also allow for an innovative form of conjecturing by finding connections between seemingly unrelated data objects in different datasets that share sub-objects.

This will be based on several technologies already developed by project partners: the MathWebSearch search engine for symbolic data at **FAU**, the publication meta-data search service at **FIZ**, and the search capabilities developed for concrete mathematical data in the LMFDB (**CHA**).

This task will be led by **UL** and **FAU**, with contributions from **FIZ** and **CHA** as indicated above.

T5 System Interoperability and Computation; Sites: **UPSud** (lead), **UL**

In this task, we demonstrate the interoperability between the FAIRMat services and computational systems, by implementing a client in the SageMath system and showcasing its use in highly user-oriented applications. The minimal functionality is to allow upload/download of datasets to/from our servers directly from within SageMath. This will already allow users to interact with shared datasets at the “push of a button” – additional functionality will be added in later tasks of this work package, e.g. search and querying in **T3.4**.

Importantly, the codecs from **T2.1** are used so that the datasets on the server appear to SageMath users

in terms of their mathematical types instead of their encodings. Because SageMath is written in Python and MMT exposes a Python API, SageMath can directly reuse the reference implementation from **T3.2** so that server and client are guaranteed to use the same data model. This will allow seamless and immediate computation with *FAIRMat* datasets, e.g., to access a dataset, perform a computation on it, and share the results as a second dataset — all in a single operation. The implementation will be based on and sustain the continuity of OpenDreamKit results, where **FAU** and **UPSud** have developed the integration of external datasets (specifically those of LMFDB) with SageMath via codecs.

In addition, this task will enable several innovative uses of the EOSC-level infrastructure. Firstly, it yields deep accessibility and deep reuse where only fragments of a dataset are accessed, computed with, or updated. Secondly, it enables the use of the *FAIRMat* sharing services as a persistent memoization layer that allows better trade-offs between data storage and on-demand re-computation, including the possibility of transparently changing between them. This would allow memoizing computationally expensive objects persistently in data stores and thus support large-scale distributed computations.

This task will be led by **UPSud**, which is one of the sites developing the SageMath system. **UL** will contribute with the DiscreteZOO package and by providing the needed server-side interfaces.

T6 FAIR Modelica; Sites: **CAE** (lead), **EMS**

Similarly to **T3.5** for the computational system SageMath, this task will develop client side functionality, in this case for the modeling of cyber-physical systems based on the Modelica language.

Concretely, we pick the two very widely used Modelica implementations — OpenModelica [OM] and Dymola [DA]. The former is Open Source, while the latter is heavily used in commercial settings. We develop a client for the *FAIRMat* services as add-ons to these tools and use them to develop user-oriented added functionality.

Modelica includes the concept of replaceable components, with a requirement that they are ”plug-compatible”: it ensures that the connectivity and parameters of a model inserted are compatible with that of the component it replaces. Therefore, a common requirement for users is to identify the available models that are compatible with a given component.

Additionally, reuse of models would be greater if a modeler could determine if others have developed similar models - using the same subsystems or equations.

Identifying models with these multiple criteria requires deep findability, e.g., whether a component uses a particular equation or offers a particular connectivity.

The FAIR Modelica service will include a search feature that lets users search for model components that can be reused in the model they are currently building, with a range of appropriate and types of search.

Like for SageMath, this will enable innovative uses of an EOSC-level infrastructure. In the case of Modelica (especially in conjunction with **T4.2**), it allows using the *FAIRMat* services as a package manager for Modelica. A general package manager for Modelica libraries is not currently embedded within modeling tools. Creating such a package manager incorporating deep-findability would be a significant step beyond the state-of-the-art.

In the case of commercial Modelica libraries, which are distributed in tool-specific encrypted formats, a lesser search would be included based on the limited meta-data that can be extracted from these libraries.

This task will be led by **CAE**, which has developed or been involved in the development of multiple Modelica tools. **EMS** will contribute to it and provide academic guidance.

T7 FAIR Services for Langlands Program; Sites: **CHA** (lead), **UL**, **UPSud**

This task will develop services responding to needs in the LMFDB community. LMFDB is a large collection of more than 30 mathematical datasets totalling a few TB that have arisen in the context of the Langland’s program [LM]. These datasets will be integrated with *FAIRMat* in **T4.4**.

The services can be divided into services supporting the download of data, and computational services primarily aiming to extend and improve the content of the LMFDB.

Download services. A major challenge of the LMFDB is the size of both the databases (number of records) and the objects inside (e.g., the integers and polynomials in those records), often to the point that the sharing of datasets must be limited to reduce the server load, thus effectively reducing the FAIR levels of LMFDB despite the intention of the authors and maintainers.

We will develop a service that allows users to access specific fragments of datasets in order to overcome this limitation. Notably this requires filtering datasets based on mathematical properties, i.e., a straightforward use of SQL database queries is not always applicable. Concretely, we will run an instance of SageMath on

the server that is — via the results of **T3.5** — coupled with the shared datasets fo filter datasets, e.g., by both eliminating rows and columns. Users will be able to submit specific queries, whose results will be computed on our servers and made available as a new dataset.
While we develop this in response to the needs of users in the LMFDB community, the service itself will be applicable to any dataset shared on the *FAIRMat* servers.
Computational services. Today adding new data to the LMFDB is a task that involves (often) downloading data from the LMFDB, running code on a local computer to compute data and then uploading the data to the LMFDB. These services aim to streamline this so that contributors will not have to worry about database implementation and database communication.
The computations will be done using SageMath running on the server. The services will allow computing additional columns in existing datasets, extending existing datasets and also creating completely new datasets. For the LMFDB a prime example where services like these would be extremely helpful are computing and storing zeros of L-functions.
While developing the services with the LMFDB as a test case, this functionality will be very helpful for any other dataset.
This task will be led by **CHA**, which has been involved in LMFDB since its inception in 2008. **UPSud**, which heavily uses SageMath, will provide user feedback and contribute to parts of the SageMath implementation. **UL** will be involved to evaluate the services on other datasets.

T8 Integration with EOSC Hub; Sites: **UL** (lead), **CAE**, **CHA**, **EMS**, **FAU**, **FIZ**, **UPSud**
Our strategy in this work package is to build a stand-alone prototype service that can be easily upgraded eventually to an official EOSC service. This cross-cutting task bundles all steps needed to ensure this is possible.
In particular, we

- maintain the compatibility with existing EOSC services as listed above,
- evaluate that all APIs are easy to migrate to dedicated EOSC servers,
- develop API documentation and maintenance instructions.

This will involve coordinating with the maintainers of these EOSC service. In this context, we will also investigate if some our math-specific results can be used to enhance existing EOSC designs, e.g., for math-specific accessibility guidelines or semantic annotation.
Moreover, in order to increase the visibility of EOSC, we will (to the extent that the authors give us permission) copy the datasets we integrate in WP4 to the EOSC/EUDat infrastructure or help the respective authors to do so.
This task constitutes a general responsibility for all sites and will be coordinated by the work package leader **UL**.

It is in the nature of service development that each deliverable couples demonstrator (type DEM), software (type OTHER), and a report (type R) describing it. Correspondingly we denote these deliverables as type R. These are always developed and delivered as a package. Therefore, they are listed a single deliverable here.

Deliverables:
D3.1 (Due: 9, Type: R, Dissem.: PU, Lead: FAU) *Basic Infrastructure Report* ⇝**M2**
The running system (type DEM), source code (type OTHER), and description (type R) from **T3.1**.

D3.2 (Due: 15, Type: R, Dissem.: PU, Lead: FAU) *The* FAIRMat *Framework for Mathematical Data* ⇝**M2**
The running system (type DEM), source code (type OTHER), and description (type R) from **T3.2**.

D3.3 (Due: 18, Type: R, Dissem.: PU, Lead: UL) *User Interface Issues in* FAIRMat ⇝**M2**
The running system (type DEM), source code (type OTHER), and description (type R) from **T3.3**.

D3.4 (Due: 24, Type: R, Dissem.: PU, Lead: UL) *Indexing, Search, Finding Connections in* FAIRMat ⇝**M3**
The running system (type DEM), source code (type OTHER), and description (type R) from **T3.4**.

D3.5 (Due: 18, Type: R, Dissem.: PU, Lead: UPSud) *SageMath as a* FAIRMat *Client* ⇝**M2**
The running system (type DEM), source code (type OTHER), and description (type R) from **T3.5**.

D3.6 (Due: 30, Type: R, Dissem.: PU, Lead: CAE) *FAIR Data Services for Mathematical Models* ⇝**M3**
The running system (type DEM), source code (type OTHER), and description (type R) from **T3.6**.

D3.7 (Due: 30, Type: R, Dissem.: PU, Lead: CHA) *Supporting Langland's Program with FAIR Data Services* ⇝**M3**
The running system (type DEM), source code (type OTHER), and description (type R) from **T3.7**.

D3.8 (Due: 36, Type: R, Dissem.: PU, Lead: UPSud) *Integrating* FAIRMat *Services with the EOSC Hub* ⇝**M4**
A report about the EOSC integration from **T3.8**.

Work Package 4: Datasets								**Start:** 1
Site	**FAU**	**UPSud**	**CHA**	***UL***	**CAE**	**FIZ**	**EMS**	**all**
Effort	12	11	12	*13*	10	12	3	**73**

Objectives

The objective of this work package is the integration of a number of existing datasets into our infrastructure. The choice of datasets must satisfy the following requirements:

R1 It must pick leading large datasets in order to maximize impact of *FAIRMat* by choosing leading and.

R2 It must cover large communities of mathematicians in order to maximize visibility of *FAIRMat*.

R3 It must choose a well-balanced set of communities and areas of mathematics to evaluate our services and maximize the number of potential users.

R4 It must provide provide blueprints and tutorials for potential users from other communities.

Description

Each task chooses one dataset and imports it into our framework. All tasks are independent of each other. Therefore, the work package lead is inessential, but each task has an independent leader.

Concretely, we choose 2 large datasets from each of the three kinds of data:

- **T4.1** and **T4.2** choose datasets that heavily use symbolic data, respectively from formalized mathematics and mathematical modeling.
- **T4.4** and **T4.3** choose datasets that heavily use concrete data, focusing respectively on the LMFDB collection of datasets and various other individual datasets.
- **T4.5** and **T4.6** choose datasets that heavily use linked data, respectively from zbMATH and Wikidata.

Each task delivers the dataset itself as made available within our services and a report describing this.

T1 Symbolic data from formalized mathematics; Sites: **FAU** (lead), **UL**

The Open Archive of Formalizations (OAF) is a collection of proof assistant libraries that have already been partially converted in OMDoc as a uniform representation standard for symbolic data. This includes specifying their semantics in the format itself via meta-logics.

In this task we complete the conversion and convert them into the *FAIRMat* standard format. In particular, the representation of proofs will be challenging. Moreover, each proof assistant library is a difficult conversion task by itself because each uses a different, complex meta-logic. Therefore, we will only be pick few proof assistant libraries as examples.

This task will be led by **FAU**, which has already built the OAF.

T2 Symbolic data from mathematical modeling; Sites: **CAE** (lead), **EMS**, **FAU**

This task integrates Modelica libraries, which form a coherent group of symbolic datasets, all conforming to the Modelica language. Besides the standard library, there are hundreds of openly shared libraries (in addition to many encrypted commercial ones). Most of these are shared on GitHub and are already aggregated at `https://www.modelica.org/libraries`. That makes it easy to process all of them systematically.

To convert all libraries into the *FAIRMat* standard, we implement a converter from Modelica to MMT. In order to achieve this we require the ability to perform stages of the compilation of the Modelica models, such as resolving types and dependencies. Symbolic data structures can then be created from the components and equations in the models.

We base this on the open-source Emmo API for Modelica, which is written in Java and thus binary-compatible with MMT, so that the integration can be developed very smoothly.

This task is led by **CAE**, which has extensive experience with Modelica and in particular develops the Emmo API for it. **FAU**, which maintains the MMT API, will advise on the implementation of the converter. **EMS** will provide additional application information.

T3 Record-Encoded Data; Sites: **UL** (lead), **UPSud**

This task integrates a broad collection of individual concrete datasets, namely the ones from Figure 1.3.5. These are typically available as sets of records as in SQL databases (although they are not necessarily stored in an SQL database, especially if the record has only few fields and uses a custom encoding to save memory). Therefore, in all cases, the work consists of two steps: Firstly, we write schema theories, which

are similar to SQL schemas except for using mathematical types and additionally fixing the encoding of each field (in the sense of **T2.1**). Using DiscreteZOO we will be able to automatically generate a website stack (from a database schema to the web interface) for many datasets based on their schema theories. Secondly, we use our implementation of the codecs to systematically process the datasets.
We will make extensive use of the outreach workshops from **T5.1**, where we will invite many of the respective dataset maintainers. Especially, the writing of the schema theories will require close collaboration until, after a few case studies, we will have written good tutorials. Many of these databases are already available via SageMath, e.g., the small groups library from GAP. These will be uploaded directly via the SageMath client from **T3.5**.
This task will be led by **UL**, which has already conducted the survey of datasets. **UPSud** will contribute those datasets that are available through SageMath.

T4 LMFDB Datasets; Sites: **CHA** (lead), **UPSud**
This task integrates the datasets from LMFDB. The procedure is essentially the same as for the datasets in **T4.3**. The writing of schema theories will be done in dialog with other members of the management board of the LMFDB, and the systematic processing of the datasets is analogous to **T4.3**. This task is necessary for the full implementation of the service in **T3.7** and will also be a show case for future integration of other mathematical datasets.
This task will be led by **CHA**, whose PI is a member of the LMFDB management board and has contributed some of the LMFDB datasets. LMFDB heavily uses SageMath, and **UPSud** will contribute SageMath to simplify the work. This includes in particular the schema theory–base accessing of LMFDB datasets from SageMath.

T5 Linked Data from Publication Metadata; Sites: **FIZ** (lead)

This WP aims to create a linked dataset derived from relevant publication metadata that is collected and maintained by **FIZ** in the zbMATH and swMATH services. Publication metadata are already available in linked data form and can be easily imported. The zbMATH dataset also includes keywords, subject classifications, and symbolic formulas that allow connecting this metadata to mathematical objects in other datasets. For mathematical software packages, there are already metadata standards developed within the FORCE11 group [Smi+16], as well as viable tools to extract semantic software information from publications in a scalable way to generate linked data. In the past, this approach has resulted in the swMATH service.
However, a similar task for other mathematical research data is much more difficult due to the vagueness of the data. It is impossible to automatically extract exact semantics for symbolical data from publications. Moreover, the size of the dataset (≈ 4 Mio. publication metadata, ≈ 30 Mio. references, ≈ 1 Mio. open full-texts available for datamining) makes it challenging to enrich the data with deep semantic information.
Therefore, for creating this dataset we need to develop heuristic methods. Additionally, we need to create user interfaces that allow for data curation, facilitating the involvement of the mathematical community (starting with $> 7,000$ zbMATH reviewers) into the editorial process.
The analysis of these datasets also allow to develop standards for symbolic data within the FORCE11 community in **T5.3**.
This task will be led by **FIZ**, which creates and maintains the dataset.

T6 Linked Data from Wikidata; Sites: **FIZ** (lead), **FAU**

Wikidata is the central storage for structured data used in Wikimedia projects including Wikipedia and Wikimedia Commons. For example, Wikipedia includes factual data from Wikidata in order to display language independent facts, such as the number of inhabitants of a town being viewed, on the page. Wikidata items might contain external identifiers such as the zbMATH Work ID (https://www.wikidata.org/wiki/Property:P894) to associate non-Wikimedia information with Wikidata items, allowing for coordination with **T4.5**.
The main challenge of this task is to align the URIs in Wikidata with their semantics as described in symbolic datasets. While both datasets use identifiers, these are typically very different. In a first step, we can extract all Wikidata identifiers and use fuzzy string matching to get alignment candidates them with other knowledge bases. A preliminary case study along these lines has already been conducted by **FAU** [GC14]. In a second step, we will manually verify and extend the alignments [Mül+17] or crowd-source this process.
FIZ will lead this task. **FIZ** personnel has extensive experience with Wikidata – for instance, introduced the data-type mathematical formulæ to Wikidata [Sch+18].

FAU will contribute expertise regarding alignments.

Deliverables:

D4.1 (Due: 24, Type: R, Dissem.: PU, Lead: FAU) *Formalized Mathematics as Symbolic Data on the EOSC* ⇝**M3**

A report describing the availability of the dataset from **T4.1**.

D4.2 (Due: 24, Type: R, Dissem.: PU, Lead: CAE) *Mathematical Models as Symbolic Data on the EOSC* ⇝**M3**

A report describing the availability of the dataset from **T4.2**.

D4.3 (Due: 30, Type: R, Dissem.: PU, Lead: UL) *Record-Encoded Mathematical Data on the EOSC* ⇝**M2**

A report describing the availability of the dataset from **T4.3**.

D4.4 (Due: 30, Type: R, Dissem.: PU, Lead: CHA) *LMFDB Data on the EOSC* ⇝**M3**

A report describing the availability of the datasets from **T4.4**.

D4.5 (Due: 24, Type: R, Dissem.: PU, Lead: FIZ) *Linked Publication Data on the EOSC* ⇝**M3**

A report describing the availability of the dataset from **T4.5**.

D4.6 (Due: 30, Type: R, Dissem.: PU, Lead: FIZ) *Scientific Wikidata on the EOSC* ⇝**M3**

A report describing the availability of the dataset from **T4.6**.

Work Package 5: Community Building, Dissemination, and Outreach								**Start:** 1
Site	**FAU**	**UPSud**	**CHA**	**UL**	**CAE**	**FIZ**	***EMS***	**all**
Effort	10	3	3	8	3	9	*8*	**44**

Objectives

The objective of this work package is to *i*) further develop the community generating, curating, and utilizing mathematical data at the European scale, *ii*) foster cross-team collaboration, spread the expertise, and engage the greater community to participate in the definition and refinement of the requirements, *iii*) leverage the communities' vast investment in existing open databases by exploiting their visibility in a uniform standard and platform. This includes:

- ensuring awareness of the results in the user community,
- engaging cross communities discussions to foster scientific collaboration and conjoint development,
- spreading the expertise through workshops and training sessions,
- providing training for dataset developers how to make their datasets more visible.

Description

The activities in this work package are split into four tasks depending on the style of communication and the role of the task leader:

- **T5.1** comprises all direct, periodic, and mostly unidirectional dissemination and public communication activities that are organized through the coordinating site.
- **T5.2** conducts major workshops targeted at specific communities of researchers. These have the goal of popularizing and EOSC in general and *FAIRMat* in particular and of training researchers in the *FAIRMat* standard.
- **T5.3** conducts general outreach activities towards the broader community, including mathematicians as a whole, researchers from other disciplines, and industry.
- **T5.4** conducts outreach activities towards the general public.

This work package does not include the formal publication of our respective scientific findings, which all members organize in the usual way in the open scientific literature and announced at scientific meetings and conferences. Similarly, it goes without saying that we will publish all data, software, source code, scientific papers, and documents openly. Open access to all publications resulting from the project will be ensured.

Moreover, we will also participate in the concertation activities, consultations, and related and events of related European E-Infrastructure projects, in particular those for the EOSC.

The work package is nominally led by **EMS**, but the primary responsibility is with the respective task leaders.

T1 Central Dissemination Activities; Sites: **FAU** (lead), **CAE**, **CHA**, **EMS**, **FIZ**, **UL**, **UPSud**

Central activities include press releases at major project milestones, maintenance of the project web-site (including visitor analysis and monitoring tools), project-wide scientific and technical publications, outreach activities (seminars, keynote talks, media interviews), promotion through social media (e.g. Twitter, Facebook, LinkedIn), the creation of advertisement materials such as flyers, posters, and electronic feeds as well as their distribution. We will use standard community building technology such as mailing lists, Wikis and Forums, to ensure dissemination and engagement of the community to support this.

This task will be led by the coordinating site. All other sites will contribute, e.g., in the form of articles for the website, specific press releases, or social media posts.

T2 Community Building: Training Workshops; Sites: **UL** (lead), **CAE**, **CHA**, **EMS**, **FAU**, **FIZ**, **UPSud**

We will organize a number of community building workshops throughout the project. This includes in particular

- announcement-style workshops open to all researchers at major conferences, in particular the European Congress of Mathematicians in 2020 and the International Congress of Mathematicians in 2022,
- Data Carpentry [DC] training sessions on a Open Scientific Data,
- two major events, which we tentatively call "Summer of Math Data" and which will be the flagship meetings centered on *FAIRMat*.

The European Congress of Mathematicians is the biggest European meeting of mathematicians and is co-organized in 2020 by the **EMS** and **UL** sites. In the summer of 2022, we will organize a major workshop at the 4-annual International Congress of Mathematicians, where we will officially release the final results of *FAIRMat*.

The Summer of Math Data events will take place in 2020 and 2021 (hosted by **UL**) and will bring together *i*) providers of mathematical datasets, *ii*) service developers, and *iii*) existing and potential users of these services in order to popularize *FAIRMat* and Open Data.

They will be organized as multi-week retreats (organized by **UL**) on the Slovenian coast. They will feature a series of partially overlapping research visits, anchored by short workshops open to all mathematicians and subsume consecutive meetings of multiple kinds:

- invitation-only parts that bring together a core group of people for coding and research sprints,
- public multi-day workshops on FAIR mathematics organized by the *FAIRMat* members,
- existing conferences, for which we will submit hosting bids, in particular the Conference on Intelligent Computer Mathematics (CICM; where **FAU** PIs are in the steering committee), the International Congress on Mathematical Software (ICMS; the *FAIRMat* coordinator is in the steering committee), and the International Congress of Mathematicians (ICM; together with the IMKT)
- the annual internal project meeting of all *FAIRMat* sites.

They will also implement and evaluate key aspects of *FAIRMat*, including the user interface and other heavily user-oriented components. They provide an efficient way to communicate with experts outside of the *FAIRMat*. In particular, several external researchers will be invited to attend the workshops, with expenses partially covered by *FAIRMat*, especially the dataset providers with whom we need to collaborate in Task **T4.3**). A typical collaboration would consist in inviting a few dataset providers, teach them how to make their datasets available in the *FAIRMat* data standard, and gather feedback from them on the *FAIRMat* services.

This task will be led by **UL** and will tie in particularly with **T4.3**. All other sites contribute.

T3 Outreach to Researchers and Industry; Sites: **EMS** (lead), **CAE**, **CHA**, **FAU**, **FIZ**, **UL**, **UPSud**

In this task, we conduct general outreach activities targeted at researchers and industry practitioners. These will take the form of official communications, workshops at major mathematical meetings, and targeted communications to specific communities. This includes transdisciplinary outreach to related fields that involve mathematical data, in particular computer science, physics, life sciences, and engineering. We will build on existing research communities connected with the partners, such as the more than 2,000 individual EMS members or more than 7,000 zbMATH reviewers. APIs designed for specific community needs will propel the adaptation of the services further. As a model serves, e.g., the open zbMATH API for the MathOverflow community site [MOa], which is currently the largest online community in mathematics ($\approx 90K$ registered users). A lean API facilitates the integration of references into the discussed questions there, and allows vice versa for the seamless interlinking of the literature with ongoing research. Similar advantages can be expected from availability of research data. Beyond mathematics, this task will be synergistically supported by the system APIs to RADAR research data hosting service provided by **FIZ**: researchers from other domains can host their research data on RADAR and integrate mathematical aspects into the *FAIRMat* framework. The expertise developed in *FAIRMat* on standardization and dissemination of symbolic data resources will fed into FORCE11 community standards like previously for software standardization.

This task will be led by **EMS**, which has the institutional support and recognition to conduct formal outreach activities. All other sites will contribute.

T4 Public Outreach via Wikipedia; Sites: **FIZ** (lead), **FAU**

Wikipedia is a major way that non-mathematicians learn about and interact with mathematical knowledge. Therefore, in order to make the public aware of scientific data collections and their benefits, we develop an integration of our datasets with Wikipedia.

Concretely, this will take the form of a Wikidata interface for the *FAIRMat* services. We compute an abstraction of all the data of *FAIRMat* that can be represented as linked data. Using the same alignments as developed in **T4.6** (but this time in the opposite direction), we will export this linked data into Wikidata and thus Wikipedia. Simultaneously, we can take advantage of the existing Wikidata infrastructure and community to provide feedback and enhancements. A specific advantage of our approach is that the semantic tool and alignments developed within the project allow to address the language barrier which often prohibits the dissemination of mathematical information beyond the field. This will yield an impact of *FAIRMat* much beyond the mathematical, and even the research community into the public discourse.

This task will be led by **FIZ**, which also leads the corresponding task **T4.6**.

The first three tasks are continuous activities and do not have natural due dates for deliverables. Therefore, we will provide two reports each at Month 18 resp. 36 that summarize the activities and their impacts; in each case, the second report will be an update to the first one.

Deliverables:

D5.1 (Due: 18, Type: R, Dissem.: PU, Lead: FAU) *Central Dissemination (intermediate report)* ⇝**M1**
The first report about the activities of **T5.1**.

D5.2 (Due: 36, Type: R, Dissem.: PU, Lead: FAU) *Central Dissemination final report* ⇝**M1**
The final report about the activities of **T5.1**.

D5.3 (Due: 18, Type: R, Dissem.: PU, Lead: UL) *Community Building (intermediate report)* ⇝**M1**
The first report about the activities of **T5.2**.

D5.4 (Due: 36, Type: R, Dissem.: PU, Lead: UL) *Community Building (final report)* ⇝**M1**
The final report about the activities of **T5.2**.

D5.5 (Due: 18, Type: R, Dissem.: PU, Lead: EMS) *General Outreach (intermediate report)* ⇝**M1**
The first report about the activities of **T5.3**.

D5.6 (Due: 36, Type: R, Dissem.: PU, Lead: EMS) *General Outreach: final report* ⇝**M1**
The final report about the activities of **T5.3**.

D5.7 (Due: 24, Type: R, Dissem.: PU, Lead: FIZ) *Public Outreach via Wikipedia* ⇝**M1**
A report about the Wikipedia integration developed in **T5.4**.

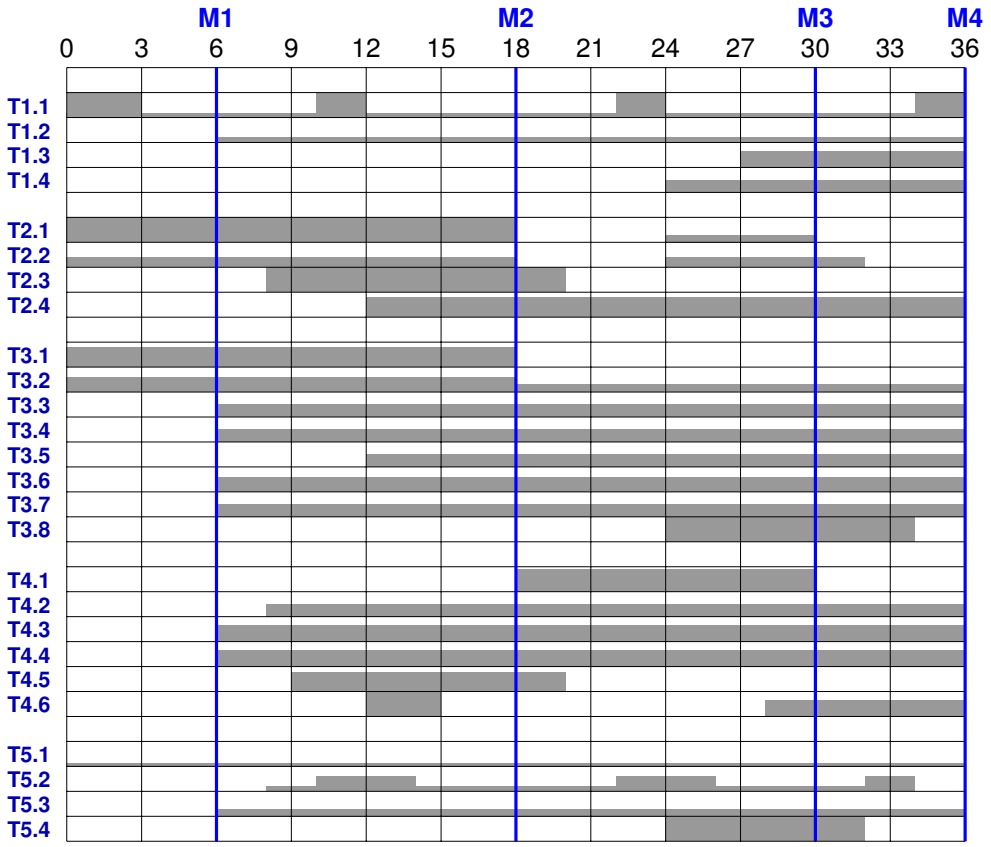

Figure 3.1.2: Gantt Chart: Overview Work Package Activities – *Bars shown at reduced height (e.g. 50%) indicate reduced intensity during that work phase (e.g. to 50%).*

3.1.3 List of Deliverables

#	Deliverable name	WP	Lead	Type	Level	Due
D1.1	Basic project infrastructure (websites, wikis, issue trackers, mailing lists, repositories)	WP1	**FAU**	DEC	PU	2
D1.2	Data Management Plan	WP1	**FAU**	R	PU	6
D3.1	Basic Infrastructure Report	WP3	**FAU**	R	PU	9
D2.1	Datatypes and Codecs	WP2	**FAU**	R	PU	12
D2.2	Metadata Ontology	WP2	**FIZ**	DEM	PU	12
D1.3	Internal Progress Report year 1, including risk management and quality assurance plan	WP1	**FAU**	R	CO	15
D3.2	The *FAIRMat* Framework for Mathematical Data	WP3	**FAU**	R	PU	15
D2.3	Legal Issues for Math Data/Services in EOSC	WP2	**EMS**	DEM	PU	18
D3.3	User Interface Issues in *FAIRMat*	WP3	**UL**	R	PU	18
D3.5	SageMath as a *FAIRMat* Client	WP3	**UPSud**	R	PU	18
D5.1	Central Dissemination (intermediate report)	WP5	**FAU**	R	PU	18
D5.3	Community Building (intermediate report)	WP5	**UL**	R	PU	18
D5.5	General Outreach (intermediate report)	WP5	**EMS**	R	PU	18
D1.4	Sustainability Plan	WP1	**EMS**	R	CO	24
D1.5	Internal Progress Report year 2	WP1	**FAU**	R	CO	24
D3.4	Indexing, Search, Finding Connections in *FAIRMat*	WP3	**UL**	R	PU	24
D4.1	Formalized Mathematics as Symbolic Data on the EOSC	WP4	**FAU**	R	PU	24
D4.2	Mathematical Models as Symbolic Data on the EOSC	WP4	**CAE**	R	PU	24
D4.5	Linked Publication Data on the EOSC	WP4	**FIZ**	R	PU	24
D5.7	Public Outreach via Wikipedia	WP5	**FIZ**	R	PU	24
D2.4	Standard Math Data on EOSC	WP2	**FAU**	DEM	PU	30
D3.6	FAIR Data Services for Mathematical Models	WP3	**CAE**	R	PU	30
D3.7	Supporting Langland's Program with FAIR Data Services	WP3	**CHA**	R	PU	30
D4.3	Record-Encoded Mathematical Data on the EOSC	WP4	**UL**	R	PU	30
D4.4	LMFDB Data on the EOSC	WP4	**CHA**	R	PU	30
D4.6	Scientific Wikidata on the EOSC	WP4	**FIZ**	R	PU	30
D1.6	Final Achievement Report	WP1	**FAU**	R	CO	36
D3.8	Integrating *FAIRMat* Services with the EOSC Hub	WP3	**UPSud**	R	PU	36
D5.2	Central Dissemination final report	WP5	**FAU**	R	PU	36
D5.4	Community Building (final report)	WP5	**UL**	R	PU	36
D5.6	General Outreach: final report	WP5	**EMS**	R	PU	36

3.2 Management Structure, Milestones, and Procedures

3.2.1 Management

3.2.2 Overview of the Organisational Structure

The following bodies will form the organizational structure of the *FAIRMat* project: Coordination Team (CT), Steering Committee (SC), Advisory Board (AB), and End User Group (EUG). The organizational structure, shown in the Figure 3.2.1, has been designed to enable efficient coordination of the project — the development and evaluation of innovative math-data-centric services involving both academic actors and industrial stakeholders.

The structure is a simplified form of the one employed in the OpenDreamKit Project, where it has worked very well. As the proposed *FAIRMat* consortium is much smaller than OpenDreamKit, we have *i*) restricted the Coordination Team (CT) to the coordinator, deputy, and project manager (they are all at **FAU** and can therefore meet easily), *ii*) dropped the quality review board (this role is taken up by the CT), and *iii*) simplified the end user board to make it more flexible.

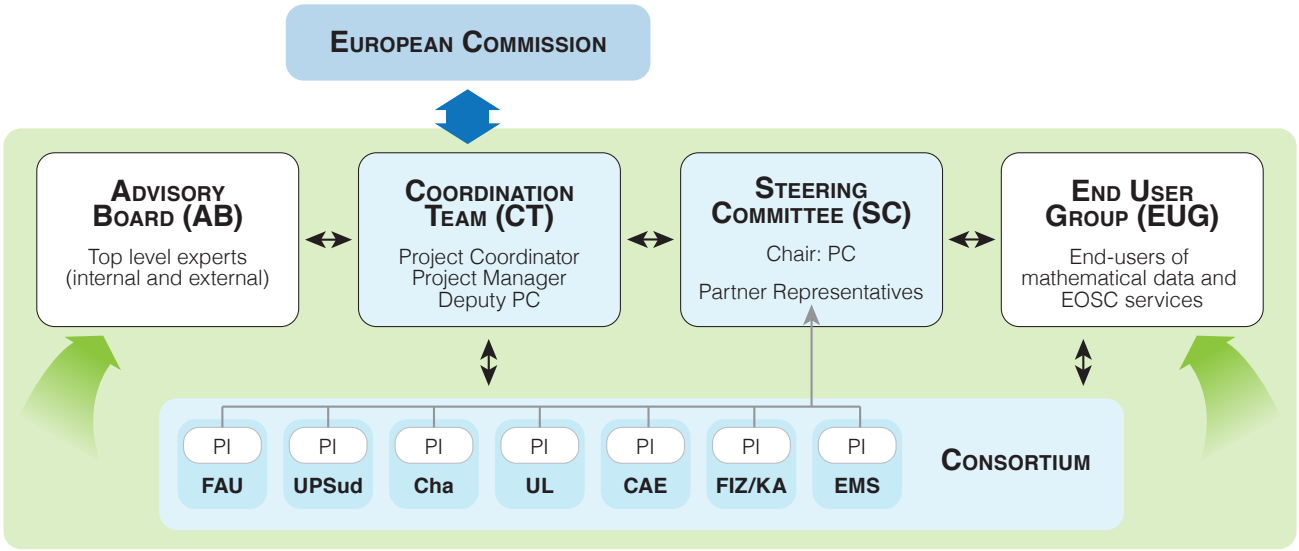

Figure 3.2.1: Management structure

We have designed the management structure and procedures to deal in a flexible manner with the following challenges:

1. to integrate all consortium members and to mobilise their expertise, knowledge and networks at every stage of the project;
2. to give the maximum attention to the end-users needs and requirements;
3. to continuously involve expertise and knowledge of relevant stakeholders and their networks, and
4. to efficiently coordinate the project implementation in a collaborative environment and ensure its sustainability.

The coordinator acts as an intermediary between the Partners and the European Commission. The coordinator will oversee the project planning, monitor that execution is carried out in time and that the objectives are achieved and closely interact with the project officer for project monitoring and delivery of the performance indicators. The Project Manager will ensure efficient day-to-day management of the project, reporting, feedback to partners on administrative, financial and legal issues, tracking of resource allocation and consumption, and communication inside and outside the consortium.

The resources of all partners will be mobilised by decentralisation of responsibilities through the assignment of leadership for work packages. Clear distribution of tasks, efficient decision making mechanisms and a sound financial management will safeguard the achievement of the project's objectives.

3.2.3 Project Coordinator and Coordinating Team (CT)

Members The project will be coordinated by the Friedrich Alexander Universität Erlangen-Nürnberg (**FAU**), represented by Prof. Michael Kohlhase (**Project Coordinator**), who has extensive experience in successfully managing research projects on the main *FAIRMat* topics. PD Dr. Florian Rabe will be employed on the project at **FAU** and serve as deputy coordinator. Kohlhase and Rabe have 15 years of collaboration experience and are the main authors of this proposal. The Project Coordinator will be assisted by a part-time (50%) **Project Manager**, who will be hired for this project and be part of the Project Coordinator's team at **FAU**. Additional feedback and expertise will be brought by FAU's Office for EU Grants and Contracts especially with regard to financial and legal matters. Together, they form the **Coordinating Team**.

Responsibilities The CT is an executive body in charge of the project implementation and monitoring. It takes operational decisions necessary for the smooth execution of the project.

Tasks

1. Monitoring the timely execution of the tasks and achievement of the objectives;
2. Preparation of scientific and financial progress reports;
3. Controlling Work Package progress by assessing it through technical reports developed by the partners;
4. Making proposals to the SC of re-allocation of tasks, resources and financial needs for the fulfilment of the work plan;
5. Preparing the drafts and validating the project deliverables to be submitted to the Commission;

Meetings: The CT can meet any time and at least twice a week. They will keep the SC updated on all executive decisions and have a bi-monthly teleconference with the SC for strategy discussions. If necessary, extra meetings will be arranged.

3.2.4 Steering Committee (SC)

Members The SC is chaired by the Project Coordinator and includes one principal investigator from each partner organization.

Responsibilities The SC is the formal decision-making body in charge of the strategic orientation of the project. It takes decisions on scientific directions, re-allocation of resources, consortium changes, and intellectual property rights.

Meetings Every 6 months. If necessary, extra-meetings will be arranged. Written minutes of each meeting will be produced, which shall be the formal record of all decisions taken. A procedure for comment and acceptance is proposed.

Voting procedure: The SC shall not deliberate until a quorum of three-fourth (3/4) of all Members are present (possibly through video-conference) or represented. Each Member shall have one vote. The SC will work on consensual decisions as much as possible and resort to voting only if unavoidable. Voting decisions shall be taken by a majority of two-thirds (2/3) of votes with quorum two-thirds (2/3) of the whole set of members. Exceptional decisions (large changes to the budget ($\geq$ 100k euros), evolution to the consortium, firing the coordinator, resolving ambiguity about whether something is a hard question) shall be taken by a majority of three-fourth (3/4) of votes with quorum three-fourth (3/4) of the whole set of members. Votes can be electronic, every member can hold a proxy vote for at most one member who cannot participate.

3.2.5 Advisory board (AB)

Members The AB consists of top level experts from partner and external organizations, including both experts from the project scientific area, and experts on legal and social matters. Potential candidates include

1. Prof. Neil Sloane, founder of the Online Encyclopedia of Integer Sequences (OEIS),
2. Prof. Bettina Eick, heavy contributor (computational packages and data) to the GAP computer algebra system,
3. Prof. Bruno Buchberger, founder of the Research Institute of Symbolic Computation (RISC Linz),
4. Prof. Ursula Martin, Mathematical Social Machines,
5. a math-affine member of a suitable DIN or ISO standardization committee,
6. Prof. Stephen M. Watt, Dean of Mathematics at University of Waterloo, co-founder of Maplesoft, co-founder of the International Mathematical Knowledge Trust (IMKT),

7. Dr. Jukka Kohonen, has published the only mathematical EOSC dataset so far,
8. Dr. Paul-Olivier Dehaye, social entrepreneur, mathematician, and data protection activist.

Some of these have been contacted and have already expressed their willingness to serve in the AB.

Responsibilities to give an independent opinion on scientific and innovation matters, in order to guaranty quality implementation of the project, efficient innovation management and project sustainability.

Meetings The AB will be invited to at least all project review meetings. Further meetings are possible at the discretion of the AB or request of the SC.

3.2.6 End User Group (EUG)

Members The EUG consists of active and visible users of the mathematical data and services; internal and external to the consortium, from different disciplines and both from academic and industrial sector. They are actively involved into the project execution, and work in close interaction with the consortium (specifically the developers) and reports to the SC – the CT monitors and the project manager tracks the proposed actions). Potential candidates include

1. Dr. Martin Otter and Prof. Peter Fritzson, chair-people of the Modelica Association,
2. Prof. Anne Frühbis-Krüger, computer algebra & data specialist,
3. Dr. David Carlisle (Numerical Algorithms Group (NAG), Oxford), LaTeX guru, Editor of the MathML Recommendations, XSLT hacker,
4. Andre Gaul, developer of PaperHive, an open annotation and review platform for academic publications,
5. Prof Vivian Pons, SageMath developer/ user, educator, and academic diversity advocate,
6. Prof. Jörg Arndt (TH Simon Ohm, Nürnberg) OEIS Editor,
7. Prof. John Cremona and Dr. David Farmer, senior members of the L-functions and Modular Forms Database (LMFDB),
8. Dietmar Winkler (University of South-Eastern Norway) works on modeling of hydroelectric power systems using Modelica, has developed commercial and open-source libraries and is a member of the Modelica Association.

Several of these have already expressed their willingness to serve in the EUG.

Responsibilities The EUG is the main actor of the innovation management within the consortium, as they have a deep understanding of both market and technical problems, and awareness of opportunities. The EUG also plays a main role in ensuring the sustainability of mathematical data and services on the EOSC.

Meetings The EUG will have regular virtual meetings and are invited to the development workshops organized in Task **T5.2**).

3.2.7 Milestones

By nature and design the project consists of a jointly designed and developed framework (in WP2) and a large number of loosely coupled tasks for the services (WP3), which are evaluated on joint case studies (WP4). We have therefore chosen to schedule the first milestone (**M1**) to track general progress on the project. In milestone **M2**, we synchronize the first fully functional version of the *FAIRMat* data standard and showcase first prototypes of the services, which will be essentially completed in milestone **M3**, which also marks initial evaluation activities. Milestone **M4** wraps up the project and marks the completion of all deliverables.

The milestones have been scheduled before the project yearly meetings in early summer, where they can be discussed in detail, tracking the progress in each work package through status reports on the tasks and deliverables and take corrective measures, where necessary, and critical decisions regarding further plans. The later milestones coincide as well with the formal project reviews for demonstration, assessment and discussion with the reviewers. We envisage that this setup will give the project the vital coherence in spite of the broad interdisciplinary mix of various backgrounds of the participants.

General Milestones

1. **Milestone M1 (Month6) Startup & Requirements Analysis** We will have carried out the requirements study and started community building activities.

2. **Milestone M2 (Month18) Prototypical Implementations** We will have constructed first version of the framework, data standards, and functional implementations of the services.

3. **Milestone M3 (Month30) Framework, Standards, Services, & Evaluation** We will have released final versions of all *FAIRMat* framework, data standards, and started the evaluation.

4. **Milestone M4 (Month36) Completed the Project** Evaluation and Documentation completed.

A tabulation of the milestones, which work packages are involved, and a means of verification can be seen in Table 3.2.1.

#	Name	WPs*/Deliverables involved	Mo	Means of Verif.
M1	*Startup & Requirements Analysis*	D1.1 D1.2 D5.1 D5.2 D5.3 D5.4 D5.5 D5.6 D5.7	6	Completed all corresponding deliverables.
M2	*Prototypical Implementations*	D1.3 D1.4 D2.1 D2.2 D2.3 D3.1 D3.2 D3.3 D3.5 D4.3	18	Completed all corresponding deliverables.
M3	*Framework, Standards, Services, & Evaluation*	D1.5 D2.4 D3.4 D3.6 D3.7 D4.1 D4.2 D4.4 D4.5 D4.6	30	Completed all corresponding deliverables.
M4	*Completed the Project*	D1.6 D3.8	36	Completed all corresponding deliverables.

Table 3.2.1: Milestones, Deliverables, and Verification (*WP is first number in deliverable label)

3.2.8 Project management tools and procedures

Project partners and management bodies will communicate through the GitLab DevOps Manager. GitLab [GL] integrates software development, versioned storage and project management facilities, is an open source software system itself, and using such a system for project management has proven successful in the OpenDreamKit project (that used GitHub [GH], whose functionality is subsumed by the open-source alternative GitLab). The Project Manager will set up the necessary GitLab infrastructure at the beginning of the project.

WP leaders will monitor progress of participants of their WP at least monthly, and participants will inform their WP leaders when problems are encountered. Major problems will be discussed in (teleconference) meetings with the CT. Each WP leader will be free to organise extra meetings with WP partners, if necessary. Scientific and financial progress reports will be collected, assembled and transmitted to the Project Coordinator by the WP leaders through the web platform. On basis of the Progress Reports, the Coordination Team will monitor progress of the project, identify bottlenecks and find solutions for these problems. Where needed, adaptations to the project plan will be made, with the aim of ensuring the delivery of the project results as agreed with the EC. Major adaptations need to be approved by the Steering Committee.

Finally, the EUG, working in close cooperation with the CT, will ensure efficient innovation management. They will carefully monitor new opportunities in order to give, if necessary, new directions to the project. For legal aspects, they will have a feedback from legal officers from the Coordinator's FAU's Office for EU Grants and Contracts, specialised in Intellectual Property.

Our management structure and procedures will ensure that our network of seven partners from both academic and industrial sectors is focused at achieving the promised deliverables, efficiently managing the innovation process and largely opening the EOSC math services to its final users. The seven partners will sign a Consortium Agreement, in which operational rules and decision making procedures will be laid down.

3.2.9 Risk management

The risk in the project execution as planned is carefully assessed and managed. We base our plans on long standing experience, and we bring together the world's experts in the relevant tools and techniques.

Our open source approach means that all our code and outputs are open and visible to anybody at sites like Github and GitLab throughout the project. This results in risk reduction: where our design decision or technical approaches are controversial, this will be detected early by those users, giving the consortium useful feedback to consider.

The CT will, with support from the SC and EUC, create a Risk Management Plan D1.3 as part of the Management Work Package, which will be reviewed annually. An initial risk assessment appears as figure 3.2.2.

Risk	with/without mitigation	Planned Mitigation measures
Recruitment of highly qualified staff	High/Medium	Great care was taken identifying pool of candidates to hire from, and coordinating with currently running projects to rehire personnel with strong track record. Typically, we will try to rehire personnel from the OpenDreamKit project that terminates in August 2019.
Different groups not forming effective team	Medium/Low	Long track record of working collaboratively on code across multiple sites; aggressive planning of project meetings, work-shops and one-to-one partner visits to facilitate most effective teamwork, combining face-to-face time at one site with remote collaboration.
Implementing services that do not match the needs of users	High/Low	Most members of the consortium are themselves end-users with a diverse range of needs and points of views; hence the design of the proposal and the governance of the project is naturally steered by demand; besides, because we provide a toolkit, users have the flexibility to adapt the infrastructure to their needs.
Reliance on external software components	Medium/Low	The non trivial software components *FAIRMat* relies on are open source. Most are very mature and supported by an active community, which offers strong long run guarantees. The components could be replaced by alternatives, or worst comes to worst, taken over by the participants.
Failure to produce a coherent service	Medium/Low	If the coordination between the framework group and the service developers is weak, this may result in incoherent services that fail to attract users. This is countered open standardization of the framework and by regular development workshops, where the stakeholders of a service meet. Moreover, the services we envision already exist at least in open prototype form, which have (only) scaled and ported to the uniform framework (by their developers).

Figure 3.2.2: Initial Risk Assessment

3.3 Consortium as a Whole

3.3.1 Overview

The consortium brings together groups of researchers and developers from different fields of expertise;

G1. Lead or core developers of a cross-section of the major mathematical databases on combinatory objects (**UL**, **CHA**).
G2. Experts in mathematical knowledge management (**FAU**, **UPSud**, **FIZ**).
G3. Experts in connecting computer algebra systems with mathematical data (**UPSud**, **UL**).
G4. Leaders in software and theory for mathematical modeling and simulation **CAE**, **EMS**; the latter via its President Volker Mehrmann, whose research focus this is.
G5. One of the two major mathematical information systems (**FIZ** provides the zbMATH and swMATH information systems for mathematical literature and software).
G6. Two companies from the industrial sector (**CAE**, and **FIZ**).
G7. The top-level professional societies for pure and applied mathematics (**EMS**; but Volker Mehrmann has also been president of the GAMM, the top-level society for applied mathematics).

There are many existing points of contact between these groups and communities, although many of them are also new to one another. This, together with the fact that each community is internally collaborative and part of the broader free software community gives us confidence in their ability to work together.

3.3.2 Roles and Resources

The exact role of each partner in each work package is defined in 3.1.2, but in general terms we have the following situation (see also Figure 3.3.1):

- Groups **G1**, **G2**, and **G3** (from the list above) will collaborate on the design and implementation of the *FAIRMat* framework, data standard, and services (WP2 and WP3).

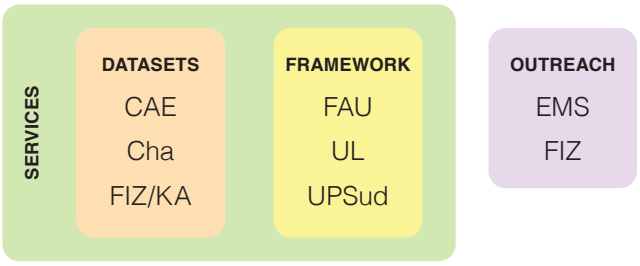

Figure 3.3.1: Overview of partner roles

- Groups **G1**, **G4**, and **G6** provide their expertise and exercise their network of clients and collaborators for evaluating the *FAIRMat* framework and infrastructure on relevant case studies (WP4).
- Finally, all participants will make use of the *FAIRMat* services, providing feedback to the developers. They will also all participate actively in dissemination activities (WP5).

The concrete researcher roles in addition to the PIs are as follows (compare the formal described of person months per site and work package in Section 3.4.1):

- **FAU**, as the coordinating site, employs 2 researchers and 0.5 admin staff:
  - PD Dr. Florian Rabe, (a senior postdoc and a main author of the proposal), will serve as deputy coordinator, lead WP2, and contribute to all other work packages. He is the designer and main developer of MMT and has extensive experience in knowledge representation languages.
  - Tom Wiesing (PhD student, later junior postdoc), will carry out the system-oriented tasks, in WP3 and WP4, which naturally expands on his current work in OpenDreamKit and as main developer of the MathHub system.
  - One administrative staff in a half-position will carry out most of the tasks in WP1 and WP5.
- **UL** will hire 1.5 researchers.
  - Dr. Katja Berčič (currently employed at **FAU** and major contributor to the proposal) will move to **UL** (full position) as a postdoc and will be the main contributor to WP3 as well as the record-oriented tasks in WP4.
  - The 0.5 FTE will be shared between Dr. Janoš Vidali, Professor Dr. Primož Potočnik, Professor Dr. Andrej Bauer, and Professor Dr. Tomaž Pisanski, who will contribute to the project from them various areas of expertise (see **UL**).
- **EMS** will hire one postdoctoral researcher with a dual focus firstly on high-level and legal aspects of the standard development and community outreach, and secondly on Modelica-related research.
- **FIZ** will fund one post-doctoral researcher with a focus on linked data, mathematical ontology, and Wikidata. This will presumably be Dr. Moritz Schubotz, who will join FIZ already in April 2019 and has extensive previous experience on these aspects.
- **CHA** will hire the PI for one year and a post-doctoral researcher for two-years postdoc. They will focus on record-encoded datasets in general and LMFDB in particular.
- **CAE** will hire one industrial software engineer with expertise in language parsing and compiler construction. Their focus will be on implementation of the Modelica-related services and tools.
- **UPSud** will hire one research software engineer, who will focus on SageMath-related implementation work.

3.3.3 Previous Collaborations

The consortium members have collaborated extensively in the past:

CAE and **EMS**: **CAE** has collaborated with Volker Mehrmann (**EMS**) as part of the MANDAE (Modelling and Advanced Numerics for Differential Algebraic Equations) consortium, a project aiming to build an Innovative Training Network and utilise Modelica as a source of mathematical models.

CHA, **FAU**, and **UPSud** collaborated through the LMFDB.

EMS and **FAU**: Volker Mehrmann (**EMS**) has collaborated on several publications with coauthors from **FAU**.

EMS and **FIZ** are partners in the the EuDML Initiative, an association without legal personality, formed by a network of universities, public research institutions, publishers, scientific databases administrators and leading technology providers, and further subjects, dedicated to producing quality scientific information in mathematics. The purpose of the Initiative is to provide a Digital Mathematics Library (DML) for the worldwide scientific

	FAU	UPSud	CHA	UL	CAE	FIZ	EMS
FAU		★★★★●@☺	@	★@		★★★●@	★★★○
UPSud	★★★★●@☺		@	○○@			
CHA	@	@					
UL	★@	○○@				@	○
CAE							●
FIZ	★★★●@			@			●○@
EMS	★★★○			○	●	●○@	
joint	★≙ publication, ●≙ project, ○≙ organization, @≙ software/resource dev, ☺≙ supervision						

Table 3.3.1: Previous Collaboration between *FAIRMat* members

community as a public service.

EMS and **FIZ** (together with the Heidelberg Academy) have been cooperating on editing zbMATH for many years, the world's most comprehensive and longest-running abstracting and reviewing service in pure and applied mathematics.

EMS and **FIZ**: Olaf Teschke (**FIZ**) is a member of **EMS**, former editor of the EMS Newsletter, and current Vice-Chair of the EMS Committee on Publication and Electronic Dissemination.

EMS and **UL**: Slovenia will host upcoming European Congress of Mathematics in 2020

FAU and **FIZ**: The MathSearch project is a longterm collaboration between the KWARC group (originally at Jacobs University, Bremen, now **FAU**) and **FIZ**.

FAU and **UPSud** are both partners in the OpenDreamKit project (2015-2019), which brings together a range of projects and associate software to create and strengthen virtual research environments, especially for mathematics. Incidentally the *FAIRMat* deputy coordinator Florian Rabe held a joint appointment between **FAU** and **UPSud** in OpenDreamKit.

FAU and **UL**: The DiscreteZOO project was originally started at **UL** and is now a collaboration between **FAU** and **UL**.

FIZ and **UL** collaborated on the application of graph-theoretic techniques to the analysis of mathematical content and its visualization in an experiment in 2010-2012 (at **UL**, Vladimir Batagelj and Tomaž Pisanski).

UPSud and **UL** Samuel Lelièvre (**UPSud**) and Katja Berčič (**UL**) worked together on the organizing committee of both Software Tools for Mathematics workshops (Morelia, Mexico and Koper, Slovenia). These workshops are aimed at students and researchers.

UPSud and **UL** are dedicating significant effort to connect computer algebra systems (particularly SageMath with mathematical databases): LMFDB (**UPSud** through OpenDreamKit) and DiscreteZOO (Janoš Vidali at **UL**). They will be able to join forces through *FAIRMat*.

3.4 Resources to be Committed

3.4.1 Person months

An overview over the distribution of person months is displayed in table 3.4.1. An explanation of the profiles of the respective positions is given in Section 3.3.2.

3.4.2 Other direct costs

The outreach aspect of *FAIRMat* requires substantial investments into hosting and attending workshops. In particular, in **T5.2**, we will invite dataset providers to workshops hosted by **UL**. Consequently, the other direct costs at **UL** will exceed 15% of its personnel costs. These are tabled below in detail. All other sites have other direct costs below 15% of personnel costs so that (as per the official instructions) no detailed tabulation is necessary.

WP	Title	FAU	UPSud	CHA	UL	CAE	FIZ	EMS	total
WP1	Management	*18*	1	1	1	1	1	10	**33**
WP2	Framework	*22*	5	5	5	5	11	12	**65**
WP3	Services	28	16	15	*27*	17	3	3	**109**
WP4	Datasets	12	11	12	*13*	10	12	3	**73**
WP5	Dissemination and Community Organization	10	3	3	8	3	9	*8*	**44**
totals		**90**	**36**	**36**	**54**	**36**	**36**	**36**	**324**

Efforts in PM; WP lead efforts light gray italicised

Table 3.4.1: Summary of Staff Efforts

UL	Cost (€)	Justification
Publication charges	1,000	Open access publication charges
Travel	32,400	Travel (see below)
Workshop organisation	40,000	**UL** will organize two multi-week Summer of Math Data events, supporting the travel expenses and accommodation for the core group participants.
Total	73,400	

Table 3.4.2: Overview: Non-staff resources to be committed at Univerza v Ljubljani (all in €)

Travel costs explanation. We use the following guidelines for expected travel expenses: €2000 for attendance of a typical one week international conference outside Europe (including travel, subsistence, accommodation and registration), €1200 for a corresponding conference in Europe, €800 for a one-week visit of a project partner, for instance for coding sprints and one-to-one research visits. We are expect that the PI with a full time position and others working on the project will in total contribute four conferences outside Europe and nine conferences in Europe. They will also cumulatively participate in eight weeklong partner visits (including project meetings). In addition to that, Katja Berčič (**UL**) will continue the tight collaboration with **FAU**, particularly on the implementation coordination (**T3.1** and **T3.2**) with two visits of three weeks per year; estimated cost of each is €1200.

References

[AAP] D. Leemans. *Atlases of Abstract Polytopes*. URL: http://homepages.ulb.ac.be/~dleemans/resources.html (visited on 01/23/2019).

[AG] D. Leemans. *An atlas of subgroup lattices of finite almost simple groups*. URL: http://homepages.ulb.ac.be/~dleemans/atlaslat/ (visited on 01/25/2019).

[AP] M. Hartley. *Abstract Polytopes*. URL: http://www.abstract-polytopes.com/atlas/index.html (visited on 01/23/2019).

[Ben+99] P. Benner, V. Mehrmann, V. Sima, S. V. Huffel, and A. Varga. "SLICOT – A Subroutine Library in Systems and Control Theory". In: *Applied and Computational Control, Signals and Circuits* 1 (1999), pp. 499–532.

[Bera] K. Berčič. *Math Databases table*. URL: https://mathdb.mathhub.info/ (visited on 01/15/2019).

[Berb] K. Berčič. *Math Databases wiki*. URL: https://github.com/MathHubInfo/Documentation/wiki/Math-Databases (visited on 01/15/2019).

[Ber03] S. Bernstein Joseph Gelbart, ed. *An Introduction to the Langlands Program*. Birkhäuser, 2003. ISBN: 3-7643-3211-5.

[Beu+] J. D. Beule, J. Jonusas, J. Mitchell, M. Torpey, and W. Wilson. *GAP package Digraphs*. URL: https://www.gap-system.org/Packages/digraphs.html (visited on 01/25/2019).

[BSa14] C. Berg, C. Stump, and al. *FindStat: The Combinatorial Statistic Finder*. http://www.FindStat.org. [Online; accessed 31 August 2016]. 2014.

[Bus+04] S. Buswell, O. Caprotti, D. P. Carlisle, M. C. Dewar, M. Gaëtano, and M. Kohlhase. *The Open Math Standard, Version 2.0*. Tech. rep. The OpenMath Society, 2004. URL: http://www.openmath.org/standard/om20.

[BV18] K. Berčič and J. Vidali. "DiscreteZOO: Towards a Fingerprint Database of Discrete Objects". In: *Mathematical Software – ICMS 2018*. Springer International Publishing, 2018, pp. 36–44. ISBN: 978-3-319-96418-8. URL: https://link.springer.com/chapter/10.1007/978-3-319-96418-8_5.

[CB15] M. Cerinšek and V. Batagelj. "Network analysis of Zentralblatt MATH data". In: *Scientometrics* 102.1 (2015), pp. 977–1001. DOI: 10.1007/s11192-014-1419-z.

[CC] *CoCalc: Collaborative Calculation in the Cloud*. URL: https://cocalc.com (visited on 01/28/2019).

[COa] M. Conder. *Lists of regular maps, hypermaps and polytopes, trivalent symmetric graphs, and surface actions*. URL: https://www.math.auckland.ac.nz/~conder/ (visited on 01/23/2019).

[COb] P. Potočnik. *Lists of graphs of a prescribed symmetry type and valence, and some other combinatorial and algebraic structures*. URL: https://www.fmf.uni-lj.si/~potocnik/work.htm (visited on 01/23/2019).

[Con+06] M. Conder, A. Malnič, D. Marušič, and P. Potočnik. "A census of semisymmetric cubic graphs on up to 768 vertices". In: *J. Algebraic Combin.* 23.3 (2006), pp. 255–294. ISSN: 0925-9899. DOI: 10.1007/s10801-006-7397-3.

[Coq] *The Coq Proof Assistant*. URL: http://coq.inria.fr/ (visited on 07/31/2010).

[Cou14] N. R. Council. *Developing a 21st Century Global Library for Mathematics Research*. Washington, DC: The National Academies Press, 2014. DOI: 10.17226/18619.

[Cre16] J. Cremona. "The L-Functions and Modular Forms Database Project". In: *Foundations of Computational Mathematics* 16.6 (2016), pp. 1541–1553. ISSN: 1615-3383. DOI: 10.1007/s10208-016-9306-z.

[D6.518] J. Cremona, D. Müller, M. Kohlhase, M. Pfeiffer, F. Rabe, Nicolas M. Thiéry, and T. Wiesing. *Report on OpenDreamKit deliverable D6.5: GAP/SAGE/LMFDB Interface Theories and alignment in OMDoc/MMT for System Interoperability*. Deliverable D6.5. OpenDreamKit, 2018. URL: https://github.com/OpenDreamKit/OpenDreamKit/raw/master/WP6/D6.5/report-final.pdf.

[DA] *Dymola*. URL: https://www.3ds.com/products-services/catia/products/dymola (visited on 01/25/2019).

[DC] *Data Carpentry – Building Communities Teaching Universal Data Literacy*. URL: https://datacarpentry.org/ (visited on 01/25/2019).

[DG] *Digraph6 file format incompatibility*. URL: https://github.com/gap-packages/Digraphs/issues/158 (visited on 01/25/2019).

[DZa] K. Berčič. *DiscreteZOO website*. URL: https://discretezoo.xyz (visited on 01/24/2019).

[DZb] J. Vidali. *DiscreteZOO SageMath package*. URL: https://github.com/DiscreteZOO/DiscreteZOO-sage (visited on 01/24/2019).

[EBO] B. Eick, H. U. Besche, and E. O'Brien. *SmallGrp – The GAP Small Groups Library*. URL: https://www.gap-system.org/Manuals/pkg/SmallGrp-1.3/doc/chap1.html (visited on 10/13/2018).

[EET] S. Wilson and P. Potočnik. *A Census of edge-transitive tetravalent graphs*. URL: https://jan.ucc.nau.edu/~swilson/C4FullSite/index.html (visited on 01/23/2019).

[EG] *Encyclopedia of Graphs*. URL: http://atlas.gregas.eu (visited on 01/24/2019).

[EUD] *EuDML – The European Digital Mathematics Library*. URL: http://eudml.eu (visited on 08/02/2011).

[FAIR18] E. C. E. G. on FAIR Data. *Turning FAIR into reality*. 2018. DOI: 10.2777/1524.

[FL] J. Kohonen. *Lists of finite lattices (modular, semimodular, graded and geometric)*. URL: https://www.shsu.edu/mem037/Lattices.html (visited on 01/25/2019).

[GAP] T. GAP Group. *GAP – Groups, Algorithms, and Programming*. URL: http://www.gap-system.org (visited on 08/30/2016).

[GC14] D. Ginev and J. Corneli. "NNexus Reloaded". In: *Intelligent Computer Mathematics*. Ed. by S. Watt, J. Davenport, A. Sexton, P. Sojka, and J. Urban. LNCS 8543. Springer, 2014, pp. 423–426. ISBN: 978-3-319-08433-6. URL: http://arxiv.org/abs/1404.6548.

[GDML] *GDML*. URL: https://en.wikipedia.org/wiki/Global_Digital_Mathematics_Library (visited on 01/28/2019).

[GG] *GReGAS, Geometric representations and symmetries of graphs, maps and other discrete structures and applications in science*. URL: http://www.gregas.eu/ (visited on 01/24/2019).

[GGS] *Gender Gap in Science*. URL: https://gender-gap-in-science.org/ (visited on 02/24/2019).

[GH] *GitHub: Build software better, together*. URL: http://github.com (visited on 02/24/2014).

[Gin+09] D. Ginev, C. Jucovschi, S. Anca, M. Grigore, C. David, and M. Kohlhase. "An Architecture for Linguistic and Semantic Analysis on the arXMLiv Corpus". In: *Applications of Semantic Technologies (AST) Workshop at Informatik 2009*. 2009. URL: http://www.kwarc.info/lamapun/pubs/AST09_LaMaPUn+appendix.pdf.

[GL] *The first single application for the entire DevOps lifecycle – GitLab*. URL: http://gitlab.com (visited on 01/12/2019).

[Har96] J. Harrison. "HOL Light: A Tutorial Introduction". In: *Proceedings of the First International Conference on Formal Methods in Computer-Aided Design*. Springer, 1996, pp. 265–269.

[Hof+13] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. "YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia." In: *AI journal* 194 (2013), pp. 28–61. URL: https://www.sciencedirect.com/science/article/pii/S0004370212000719.

[HS13] S. Harris and A. Seaborne. *SPARQL 1.1 Query Language*. W3C Recommendation. World Wide Web Consortium (W3C), Mar. 21, 2013. URL: https://www.w3.org/TR/sparql11-query/.

[Ian+14] M. Iancu, C. Jucovschi, M. Kohlhase, and T. Wiesing. "System Description: MathHub.info". In: *Intelligent Computer Mathematics*. Ed. by S. Watt, J. Davenport, A. Sexton, P. Sojka, and J. Urban. LNCS 8543. Springer, 2014, pp. 431–434. ISBN: 978-3-319-08433-6. URL: http://kwarc.info/kohlhase/papers/cicm14-mathhub.pdf.

[JK10] C. Jucovschi and M. Kohlhase. "sTeXIDE: An Integrated Development Environment for sTeX Collections". In: *Intelligent Computer Mathematics*. Ed. by S. Autexier, J. Calmet, D. Delahaye, P. D. F. Ion, L. Rideau, R. Rioboo, and A. P. Sexton. LNAI 6167. Springer Verlag, 2010, pp. 336–344. ISBN: 3642141277. arXiv: 1005.5489v1 [cs.OH].

[KDH] *KWARC Data Host*. URL: https://datahost.kwarc.info/ (visited on 01/25/2019).

[KMP12] M. Kohlhase, B. A. Matican, and C. C. Prodescu. "MathWebSearch 0.5 – Scaling an Open Formula Search Engine". In: *Intelligent Computer Mathematics*. Ed. by J. Jeuring, J. A. Campbell, J. Carette, G. Dos Reis, P. Sojka, M. Wenzel, and V. Sorge. LNAI 7362. Berlin and Heidelberg: Springer Verlag, 2012, pp. 342–357. ISBN: 978-3-642-31373-8. URL: http://kwarc.info/kohlhase/papers/aisc12-mws.pdf.

[Koh+11] M. Kohlhase et al. "The Planetary System: Web 3.0 & Active Documents for STEM". In: *Procedia Computer Science* 4 (2011): *Special issue: Proceedings of the International Conference on Computational Science (ICCS)*. Ed. by M. Sato, S. Matsuoka, P. M. Sloot, G. D. van Albada, and J. Dongarra. Finalist at the Executable Paper Grand Challenge, pp. 598–607. DOI: 10.1016/j.procs.2011.04.063.

[Koh+17] M. Kohlhase, D. Müller, M. Pfeiffer, F. Rabe, N. Thiéry, V. Vasilyev, and T. Wiesing. "Knowledge-Based Interoperability for Mathematical Software Systems". In: *MACIS 2017*. Ed. by J. Blömer, T. Kutsia, and D. Simos. LNCS 10693. Springer Verlag, 2017, pp. 195–210. URL: https://github.com/OpenDreamKit/OpenDreamKit/blob/master/WP6/MACIS17-interop/crc.pdf.

[Koh06] M. Kohlhase. *OMDoc – An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. Springer Verlag, Aug. 2006. URL: http://omdoc.org/pubs/omdoc1.2.pdf.

[Kra+16] A. Kraft, M. Razum, J. Potthoff, A. Porzel, T. Engel, F. Lange, K. Van den Broek, and F. Furtado. "The RADAR project – a service for research data archival and publication". In: *ISPRS International Journal of Geo-Information* 5.3 (2016). ISSN: 2220-9964. DOI: 10.3390/ijgi5030028.

[KŞ06] M. Kohlhase and I. Şucan. "A Search Engine for Mathematical Formulae". In: *Proceedings of Artificial Intelligence and Symbolic Computation, AISC'2006*. Ed. by T. Ida, J. Calmet, and D. Wang. LNAI 4120. Springer Verlag, 2006, pp. 241–253. URL: http://kwarc.info/kohlhase/papers/aisc06.pdf.

[KWARC] *Knowledge Adaptation and Reasoning for Content*. URL: http://kwarc.info (visited on 05/12/2011).

[Leh+13] J. Lehmann et al. "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia". In: *Semantic Web Journal* (2013), p. 29. URL: http://www.semantic-web-journal.net/system/files/swj558.pdf.

[LK16] E. Luzhnica and M. Kohlhase. "Formula Semantification and Automated Relation Finding in the OEIS". In: *Mathematical Software - ICMS 2016 - 5th International Congress*. Ed. by G.-M. Greuel, T. Koch, P. Paule, and A. Sommese. Vol. 9725. LNCS. Springer, 2016. DOI: 10.1007/978-3-319-42432-3.

[LM] *The L-functions and Modular Forms Database*. URL: http://www.lmfdb.org (visited on 02/01/2016).

[LPK] *LAPACK — Linear Algebra PACKage*. URL: https://www.netlib.org/lapack/ (visited on 01/25/2019).

[Luz16] E. Luzhnica. "Formula Semantification and Automated Relation Finding in the OEIS". B. Sc. Thesis. Jacobs University Bremen, 2016. URL: https://github.com/eluzhnica/OEIS/doc/Enxhell_Luzhnica_BSC.pdf.

[McKa] B. McKay. *Combinatorial Data*. URL: http://users.cecs.anu.edu.au/~bdm/data/ (visited on 01/25/2019).

[McKb] B. McKay. *Graph formats*. URL: http://users.cecs.anu.edu.au/~bdm/data/formats.html (visited on 01/25/2019).

[MH] *MathHub.info: Active Mathematics*. URL: http://mathhub.info (visited on 01/28/2014).

[MLB] *MATLAB*. URL: https://www.mathworks.com/products/matlab.html (visited on 01/25/2019).

[MML310] R. Ausbrooks et al. *Mathematical Markup Language (MathML) Version 3.0*. Ed. by D. Carlisle, P. Ion, and R. Miner. 2010. URL: http://www.w3.org/TR/MathML3.

[MMTa] *MMT – Language and System for the Uniform Representation of Knowledge*. project web site. URL: https://uniformal.github.io/ (visited on 01/15/2019).

[MMTb] F. Rabe. *The MMT Language and System*. URL: https://svn.kwarc.info/repos/MMT/doc/html (visited on 10/11/2011).

[MMTc] *UniFormal/MMT – The MMT Language and System*. URL: https://github.com/UniFormal/MMT (visited on 10/24/2017).

[MOa] *MathOverflow*. URL: https://mathoverflow.net/ (visited on 01/29/2019).

[MOb] *Modelica and the Modelica Association*. URL: https://modelica.org/ (visited on 01/22/2019).

[MPS] *MUltifrontal Massively Parallel sparse direct Solver*. URL: http://mumps.enseeiht.fr/ (visited on 01/25/2019).

[Mül+17] D. Müller, T. Gauthier, C. Kaliszyk, M. Kohlhase, and F. Rabe. "Classification of Alignments between Concepts of Formal Mathematical Systems". In: *Intelligent Computer Mathematics*. Ed. by H. Geuvers, M. England, O. Hasan, F. Rabe, and O. Teschke. LNAI 10383. Springer, 2017. ISBN: 978-3-319-62074-9. DOI: 10.1007/978-3-319-62075-6.

[OAF] *OAF: An Open Archive for Formalizations*. URL: https://kwarc.info/projects/oaf/ (visited on 04/26/2018).

[OBI] *Ontology for Biomedical Investigations*. URL: http://obi-ontology.org/ (visited on 01/25/2019).

[ODAP] *OPenNDAP - Advanced Software for Remote Data Retrieval*. URL: https://opendap.org/ (visited on 01/25/2019).

[OEIS] *The On-Line Encyclopedia of Integer Sequences*. URL: http://oeis.org (visited on 05/28/2017).

[OM] *OpenModelica*. URL: https://openmodelica.org/ (visited on 01/22/2019).

[OWL09] OWL Working Group. *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation. World Wide Web Consortium (W3C), Oct. 27, 2009. URL: http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.

[PG] *PARI/GP Home*. URL: https://pari.math.u-bordeaux.fr/ (visited on 01/29/2019).

[PK11] C. C. Prodescu and M. Kohlhase. "MathWebSearch 0.5 - Open Formula Search Engine". In: *Wissens- und Erfahrungsmanagement LWA (Lernen, Wissensentdeckung und Adaptivität) Conference Proceedings*. Sept. 2011. URL: https://github.com/KWARC/mws/doc/2011/newmws/main.pdf.

[Pot09] P. Potočnik. "A list of 4-valent 2-arc-transitive graphs and finite faithful amalgams of index $(4, 2)$". In: *European J. Combin.* 30.5 (2009), pp. 1323–1336. ISSN: 0195-6698. DOI: 10.1016/j.ejc.2008.10.001.

[PSV13] P. Potočnik, P. Spiga, and G. Verret. "Cubic vertex-transitive graphs on up to 1280 vertices". In: *J. Symbolic Comput.* 50 (2013), pp. 465–477. ISSN: 0747-7171. DOI: 10.1016/j.jsc.2012.09.002.

[PSV15] P. Potočnik, P. Spiga, and G. Verret. "A census of 4-valent half-arc-transitive graphs and arc-transitive digraphs of valence two". In: *Ars Math. Contemp.* 8.1 (2015), pp. 133–148. ISSN: 1855-3966.

[RDF] World Wide Web Consortium (W3C), ed. *Resource Description Framework (RDF)*. URL: http://www.w3.org/RDF/ (visited on 10/22/2009).

[RK13] F. Rabe and M. Kohlhase. "A Scalable Module System". In: *Information & Computation* 0.230 (2013), pp. 1–54. URL: http://kwarc.info/frabe/Research/mmt.pdf.

[Roy] G. Royle. *Combinatorial Catalogues*. URL: http://staffhome.ecm.uwa.edu.au/~00013890/data.html (visited on 01/25/2019).

[Sage] T. S. Developers. *SageMath, the Sage Mathematics Software System*. URL: http://www.sagemath.org (visited on 09/30/2016).

[Sch+18] M. Schubotz, P. Scharpf, K. Dudhat, Y. Nagar, F. Hamborg, and B. Gipp. "Introducing MathQA: a Math-Aware question answering system". In: *Information Discovery and Delivery* 46.4 (2018), pp. 214–224. ISSN: 2398-6247. DOI: 10.1108/IDD-06-2018-0022.

[SCT] *SNOMED Clinical Terms*. URL: https://bioportal.bioontology.org/ontologies/SNOMEDCT (visited on 01/25/2019).

[SDB] *SageMath Databases*. URL: http://doc.sagemath.org/html/en/reference/databases/index.html (visited on 01/25/2019).

[Slo03] N. J. A. Sloane. “The On-Line Encyclopedia of Integer Sequences”. In: *Notices of the AMS* 50.8 (2003), p. 912.

[Smi+16] A. M. Smith, D. S. Katz, K. E. Niemeyer, and FORCE11 Software Citation Working Group. “Software citation principles”. In: *PeerJ Computer Science* 2 (Sept. 2016), e86. ISSN: 2376–5992. DOI: 10.7717/peerj-cs.86.

[SR14] G. Schreiber and Y. Raimond. *RDF 1.1 Primer*. W3C Working Group Note. World Wide Web Consortium (W3C), 2014. URL: http://www.w3.org/TR/rdf-primer.

[SUMO] *Suggested Upper Merged Ontology*. URL: http://www.adampease.org/OP/ (visited on 01/25/2019).

[SWM] *Mathematical Software – swMATH*. URL: http://swmath.org (visited on 09/07/2017).

[TG] B. McKay, G. Royle, and A. Hulpke. *Transitive Graphs*. URL: http://staffhome.ecm.uwa.edu.au/~00013890/remote/trans/index.html (visited on 01/23/2019).

[TH14] M. Tiller and P. Harman. “recon – Web and network friendly simulation data formats”. In: 96 (2014), pp. 1081–1093. ISSN: 1650-3740. URL: http://www.ep.liu.se/ecp/article.asp?issue=096&article=113&volume=#.

[UL] M. Malandro. *Unlabeled lattices on* $<= 15$ *nodes*. URL: https://b2share.eudat.eu/records/dbb096da4e364b5e9e37b982431f41de (visited on 01/25/2019).

[Wan] I. Wanless. *Combinatorial Data*. URL: http://users.monash.edu.au/~iwanless/data/ (visited on 01/25/2019).

[Wat+14] S. Watt, J. Davenport, A. Sexton, P. Sojka, and J. Urban, eds. *Intelligent Computer Mathematics*. LNCS 8543. Springer, 2014. ISBN: 978-3-319-08433-6.

[WD] *Wikidata:Introduction*. URL: https://wikidata.org/wiki/Wikidata:Introduction (visited on 01/25/2015).

[Wil+16] M. D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3 (2016). URL: https://doi.org/10.1038/sdata.2016.18.

[WM] *WOLFRAM MATHEMATICA The world’s definitive system for modern technical computing*. URL: https://www.wolfram.com/mathematica/ (visited on 01/24/2019).

[ZBM] *zbMATH the first resource in mathematics*. URL: http://zbmath.org (visited on 01/29/2019).

[ZEE16] M. Ziemann, Y. Eren, and A. El-Osta. “Gene name errors are widespread in the scientific literature”. In: *Genome Biology* 17.177 (2016). DOI: 10.1186/s13059-016-1044-7.

4 Members of the Consortium

4.1 Participants

4.1.1 FAU: Friedrich-Alexander Universität Erlangen/Nürnberg (DE)

Legal Entity

Friedrich Alexander Universität Erlangen/Nürnberg (FAU) is a public research university in the cities of Erlangen and Nuremberg, Germany. FAU is the second largest state university in the state Bavaria. It has 5 faculties, 23 departments/schools, 30 clinical departments, 19 autonomous departments, 656 professors, and ca 40 000 students.

FAU has a strong departments of Computer Science and Mathematics (both 25 Professors) with strong groups in scientific computing, mathematical modeling, and simulation, data management, data visuatlization, and Pattern recognition. All of these deal with mathematical data in some way and thus constitute a conducive and supportive environment for *FAIRMat*. Importantly the collaborating departments constitute a reservoir of know-how and potential user expertise the *FAIRMat* project can draw upon for evaluation and testing.

Prof. Dr. Michael Kohlhase holds the *Professorship for Knowledge Representation and Management* in the Computer Science Department. The KWARC (KnoWledge Adaptation and Reasoning for Content [KWARC]) Group headed by him specialises in knowledge management for Science, Technology, Engineering, and Mathematics (STEM), focusing on the last as a test subject. Formal logic, natural language semantics, and semantic web technology provide the foundations for the research of the group. Its group working on *FAIRMat* will be composed of the following non-exhaustive list: Prof. Dr. Michael Kohlhase, Dr. Florian Rabe, Tom Wiesing, Dennis Müller, Jonas Betzendahl and Jan Frederik Schaefer. The first two will serve as principal investigators, the third will be directly employed, and the last three will support the project with knowledge and synergistic research.

FAU will lead the project (WP1), host the majority of the man power for *FAIRMat* framework development (WP2), and contribute to the services (WP3), use cases (WP4), and dissemination/outreach (WP5). Overall, **FAU** will contribute 90 Person-Months to the *FAIRMat* project (see Table 3.4.1).

Curriculum vitae

Michael Kohlhase (leadPI, male) holds the Professorship for Knowledge Representation and Management at FAU Erlangen-Nürnberg and is an associate adjunct professor at Carnegie Mellon University.

He studied pure mathematics at the Universities of Tübingen and Bonn (1983–1989) and continued with computer science, in particular higher-order unification and automated theorem proving (Ph.D. 1994, Saarland University).

His current research interests include knowledge representation for mathematics, inference-based techniques for natural language processing, and computer-supported education. He has pursued these interests during extended visits to Carnegie Mellon University, SRI International, and the Universities of Amsterdam, Edinburgh, and Auckland.

Michael Kohlhase is recipient of the dissertation award of the Association of German Artificial Intelligence Institutes (AKI; 1995) and of a Heisenberg stipend of the German Research Council (DFG 2000-2003). He was a member of the Special Research Action 378 (Resource-Adaptive Cognitive Processes), leading projects on both automated theorem proving and computational linguistics. Michael Kohlhase was trustee of the Conference on Automated Deduction (CADE), Mathematical Knowledge Management (MKM), and the CALCULEMUS conference, he was a member of the W3C Math working group, the International Mathematical Knowledge Trust, president of the OpenMath Society and the MathML Association (MathML e.V.), and has been the general secretary of the Conference on Intelligence Computer Mathematics (CICM).

Florian Rabe (PI, male, 36 PM) is a senior post-doctoral researcher at both University Erlangen-Nuremberg, Germany, and University Paris-Sud, France. He completed his PhD in 2008 and his habilitation in 2014 and holds the venia legendi.

He has worked on the formal representation and management of mathematical knowledge for 15 years. He was a lead researcher in the LATIN project (2009-2012), which produced a highly modular and integrated library of formal languages for knowledge representation and the OAF project (2014–2019), which produced an integrated archive of symbolic mathematical data in the area of logic. He is currently a principal investigator in the OpenDreamKit, where he developed the designed many of the concepts and TRL 5-6 services on which *FAIRMat* builds, including the standard for semantics-aware mathematical datasets.

He is the creator and main developer of the MMT language and system, which are the backbone of both LATIN and OAF and a central component of OpenDreamKit. MMT has been developed for over 10 years with more than 20 contributors and currently consists of more than $100,000$ lines of Scala code.

He served in the organising committee of 5 and the program committee of 14 international conferences (3 as track chair, 1 as general chair) and has organised or served in the program committee of 24 international workshops. He is currently a member of the steering committee of the Conference on Intelligent Computer Mathematics, the secretary of the OpenMath society, the chair of the steering committe of the LFMTP workshop on logical frameworks, and a member of the IFIP WG 2.1 on Algorithmic Languages and Calculi. He has authored 89 research papers (18 in international journals), edited 4 books, and has supervised 19 undergraduate and graduate theses.

Tom Wiesing (PI, 36 PM) is a PhD candidate at University Erlangen-Nuremberg, Germany. He completed his Master of Science in Data Engineering in 2017 and his Bachelor of Science in Applied and Computational Mathematics in 2015.

He is currently employed in the OpenDreamKit project, where he mainly works on the Math-In-The-Middle approach building a scalable solution that allows integrating multiple existing mathematical software systems. Furthermore, he is the current main developer of the MathHub System, a portal for active mathematical documents and an archive for flexiformal mathematics.

He has also co-authored several papers at international, peer-reviewed conferences.

Achievements

[KMP12] M. Kohlhase, B. A. Matican, and C. C. Prodescu. "MathWebSearch 0.5 – Scaling an Open Formula Search Engine". In: *Intelligent Computer Mathematics*. Ed. by J. Jeuring, J. A. Campbell, J. Carette, G. Dos Reis, P. Sojka, M. Wenzel, and V. Sorge. LNAI 7362. Berlin and Heidelberg: Springer Verlag, 2012, pp. 342–357. ISBN: 978-3-642-31373-8. URL: http://kwarc.info/kohlhase/papers/aisc12-mws.pdf.

[Koh+11] M. Kohlhase et al. "The Planetary System: Web 3.0 & Active Documents for STEM". In: *Procedia Computer Science* 4 (2011): *Special issue: Proceedings of the International Conference on Computational Science (ICCS)*. Ed. by M. Sato, S. Matsuoka, P. M. Sloot, G. D. van Albada, and J. Dongarra. Finalist at the Executable Paper Grand Challenge, pp. 598–607. DOI: 10.1016/j.procs.2011.04.063.

[Koh06] M. Kohlhase. *OMDoc – An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. Springer Verlag, Aug. 2006. URL: http://omdoc.org/pubs/omdoc1.2.pdf.

[KR16] M. Kohlhase and F. Rabe. "QED Reloaded: Towards a Pluralistic Formal Library of Mathematical Knowledge". In: *Journal of Formalized Reasoning* 9.1 (2016), pp. 201–234. URL: http://jfr.unibo.it/article/download/4570/5733.

[RK13] F. Rabe and M. Kohlhase. "A Scalable Module System". In: *Information & Computation* 0.230 (2013), pp. 1–54. URL: http://kwarc.info/frabe/Research/mmt.pdf.

Experience

The KWARC group is the lead implementor of the OMDoc (Open Mathematical Document) format for representing mathematical knowledge [Koh06] and redeveloped its formal core in the OMDoc/MMT format [RK13]. The latter has been implemented in the MMT system [MMTb; RK13] which provides efficient implementations of the computational primitives such as type checking, flattening, and presentation at a logic/foundation-independent level. The group has developed services powered by such semantically rich representations, different paths to obtaining them, as well as platforms that integrate both aspects. *Services* include the adaptive context-sensitive presentation framework provided by the MMT API and the semantic search engine MathWebSearch[KŞ06; KMP12].

Semantic services can be integrated into the documents generated from OMDoc/MMT representations, making them into "active documents", i.e. documents that are interactive and adaptive to the user and situation. For *obtaining* rich content, the group investigates assisted manual editing [JK10] as well as automatic annotation using linguistic techniques [Gin+09]. Finally, KWARC has developed the MathHub.info portal a community-based library and knowledge management system for flexiformal libraries, which can be used for semantic publishing and eLearning [Koh+11; MH; Ian+14].

The OMDoc/MMT knowledge representation format and the MathHub.info system will an important basis for the developments Work PackagesWP2 and WP3.

Prof. Dr. Michael Kohlhase has initiated and led the CALCULEMUS! IHP-Research and Training Network and participated in the FP6 IST MoWGLI (Mathematics on the Web: Get it by Logic and Interfaces) project, the FP6 CSA Once-CS (Open Network of Centres of Excellence in Complex Systems), The FP7 EDC project WebALT (Web Advanced Learning Technologies), and has been a work package leader of the OpenDreamKit H2020 project. Finally, he has co-initiated and organize the three NTCIR community challenges for mathematics information retrieval in 2014/16/17.

- Modeling formal structures of mathematical knowledge in a web-scalable way.
- Transforming large collections of legacy scientific publications to semantically structured markup.
- Designing user interfaces for authoring and interacting with mathematical knowledge.

Infrastructure

The department of Computer Science at FAU has institutionalized a Working Group on Research Data (WGRD) for coordinating RD efforts at the university level with ultimate aim of creating a central facility for research data for FAU. Michael Kohlhase is the deputy speaker of WGRD.

The KWARC research group hosts http://MathHub.info, a portal for formalized mathematics and active mathematical documents with hosts about 10GB symbolic mathematical data (Theorem Prover Libraries, OEIS, semantic course materials and a multilingual mathematical glossary), serves it via the MMT system and a lightweight browser-based front-end, and includes services like 2D/3D theory graph visualization of the modular structure.

4.1.2 UL: UNIVERZA V LJUBLJANI (SI)

Legal Entity

The University of Ljubljana (UL) is the oldest and largest higher education and scientific research institution in Slovenia. It has 23 faculties, 3 academies, over 5.700 employees, and ca 40.100 students. The Faculty of mathematics and physics (UL FMF) employs 172 researchers who are engaged in a wide spectrum of different areas of Physics, Mathematics and Computer Science. Research and development activities are carried out in the form of projects funded by the agencies and ministries on the national level, the European Commission agencies and other companies and organisations. UL FMF currently hosts one ERC advanced grant in Mathematical Physics.

The Discrete Mathematics Group (DMG) and the Theoretical Computer Science Group (TCSG) are two of the strongest and most active research groups at UL FMF. The members of the DMG and TCSG who will be actively involved with *FAIRMat* are: Primož Potočnik, Tomaž Pisanski, Andrej Bauer and Janoš Vidali. The UL team will also be joined by Vladimir Batagelj, Professor Emeritus of UL, who has a vast experience in data analysis, and analysis of large networks and who has contributed a vast number of software tools for mathematics.

The TCSG group (lead by prof. Andrej Bauer and prof. Alexander Simpson) covers a variety of topics: type theory, homotopy type theory, formalization of mathematics, theory of programming languages, computable and constructive mathematics, and symbolic computation.

The main focus of DMG is graph theory and in particular various computational aspects thereof. Primož Potočnik is the author of several datasets of highly symmetric graphs (most prominently, the complete list of all cubic vertex-transitive graphs of order up to 1280) that are widely used in the graph theoretical research community. Producing further datasets of different combinatorial objects remains one of the main focuses of DMG. The idea for the DiscreteZOO project was conceived by Berčič, Potočnik and Vidali while Berčič was working on her PhD thesis under Potočnik's mentorship.

UL will lead the work packages WP3 and WP4, focusing on accessibility, user interface, indexing, and searching in the former and on record data in the latter.

Curriculum vitae

Primož Potočnik (leadPI, male, 4 PM) is a full professor in mathematics at the Faculty of Mathematics and Physics at the University of Ljubljana.

He studied pure mathematics at University of Ljubljana (1991-1995) and received his PhD degree in Mathematics from the University of Ljubljana in 2000. He continued his studies at the postdoctoral level at the University of Ottawa, Canada (2003), and the University of Auckland, New Zealand (2004-2005).

His research interest lies in the intersection of algebra and discrete mathematics with the emphasis on the symmetries of discrete objects. He is the author of several exhaustive databases of highly symmetric combinatorial objects, such as the census of cubic vertex-transitive graphs of order up to 1280, that are highly used in the discrete mathematics community.

Primož Potočnik has been a PI of two national research projects (3 FTE in total) and is currently holding a long-term research programme (2 FTE) financed by the Slovenian Research Agency. He was a recipient of the Fulbright Scholarship in 2003 which financed his 5 months visit to the University of Flagstaff, Arizona.

He has been actively involved in the organisation a number of major mathematical events, most notably in the organisation of the European Mathematical Congress 2020 (member of the Organising committee). He is a member of the Council of the European Mathematical Society (term 2018-2020) and is currently the Head of Department of Mathematics as well as Vice dean of the Faculty of Mathematics and Physics at the University of Ljubljana, finishing his term in September 2019.

Katja Berčič (PI, female, 36 PM) is a postdoc in the Research Group of the Professorship for Knowledge Representation and Processing, FAU Erlangen-Nürnberg.

She studied pure mathematics at the University of Ljubljana (BSc 2010, advisor Andrej Bauer). Her PhD thesis (2015, advisor Primož Potočnik) involved classifying certain classes of highly symmetric graphs as well as generating a census. Together with Potočnik and Vidali they observed a distinct lack of tools for mathematical datasets, starting the DiscreteZOO project to address it.

She pursued related interests in computer science and the industry for two years, obtaining valuable experience in working with databases and interfaces for web-based applications. A postdoc position at UNAM in Morelia was an opportunity to continue working on DiscreteZOO, as well as on another dataset of highly symmetric objects.

Katja Berčič was involved in the organisation of five conferences and workshops, two of which were Software Tools for Mathematics. She is also the main author of the survey of mathematical datasets [Berb; Bera].

Andrej Bauer (R, male, 4 PM) is a professor of computational mathematics at the Faculty of Mathematics and Physics at the University of Ljubljana.

His research is motivated by the desire to understand the connection between mathematics and computation. His areas of interest are logic, type theory, homotopy type theory, constructive and computable mathematics, principles of programming languages, proof assistants, and formalized mathematics.

He received a BSc in mathematics at the University of Ljubljana (1990–1994), and a PhD degree in Pure and Applied Logic at the School of Computer Science, Carnegie Mellon University in 2000, for which he received the Doctoral Dissertation

Award. After a postdoc at the Mittag-Leffler Instut, the Royal Swedish Academy of Sciences, he returned to Slovenia where he has worked since. In 2012 he visited the Institute for Advanced Study in Princeton, USA, where he was a member of the group of mathematicians who developed homotopy type theory. He is a coauthor of the book "Homotopy Type Theory: Univalent Foundations of Mathematics", and of the HoTT library, a large formalization of homotopy type theory in the Coq proof assistant. Since 2012 he has worked on homotopy type theory, formalization of mathematics, and implementation of new proof assistants.

Andrej Bauer was the principal investigator of Computational Effects in Computable and Formalized Mathematics (EFFMATH), Air Force Office of Scientific Research Award No. FA9550-14-1-0096 (2014–2017), and is the principal investigator of Foundations of Type Theory for Computation and Mathematics (TYPECOMA), Air Force Office of Scientific Research Award No. FA9550-17-1-0326 (2017–2020). He is a member of the Management Committee of the EU COST network CA15123, The European research network on types for programming and verification (EUTypes), and the head of the Working Group on Theoretical foundations of Type Theory.

Tomaž Pisanski (R, male, 4 PM) is a full professor in mathematics sharing his work between the Faculty of Mathematics and Physics at the University of Ljubljana, Institute of Mathematics, Physics and Mechanics, Ljubljana and University of Primorska, Koper.

He studied technical mathematics at the University of Ljubljana (1968-1972) and received his MSc in Mathematics at the University of Ljubljana (1979), MSc in Computer Science from the Penn State University (1980) and PhD degree in Mathematics from the University of Ljubljana in 1981.

His research interest lies in discrete mathematics and its applications, with the emphasis on the topological, algebraic and chemical graph theory. He has co-authored over 150 research papers, published both in pure math journals, such as J. London Math. Soc, and more applied journals such as Nature Commun. Since 2008 he has been a co-editor-in-chief of the journal Ars Mathematica Contemporanea. In 2012 he coauthored a monograph on configurations (Springer).

Tomaž Pisanski has been a PI of several national research projects and until the end of 2018 the PI of a long-term research programme financed by the Slovenian Research Agency comprising about 30 researchers. He was a recipient of the Fulbright Scholarship and a small NSF grant. He has been a Lead PI within the EUROCORES Programme EUROGIGA (project GReGAS, spanned over six European countries) of the European Science Foundation (2011-2014).

He has been actively involved in the organisation a number of major mathematical events, most notably in the organisation of the European Mathematical Congress 2020 (chair of the Organising committee). He is a member of the Publications Committee of the European Mathematical Society and is currently the Head of Department of Information Sciences and Technologies at the Faculty of Mathematics, Natural Sciences and Information Technologies at the University of Primorska. Tomaž Pisanski is a member of the Academia Europaea, Slovenian Engineering Academy and International Academy of Mathematical Chemistry. He is a Fellow of ICA and Fellow of the SDAMS.

Janoš Vidali (R, male, 6 PM) is a teaching assistant at the Faculty of Mathematics and Physics, University of Ljubljana.

On the undergraduate level, he took the Interdisciplinary study of Computer Science and Mathematics, and graduated with a thesis from cryptography, which was also his initial research area. Later, he started researching in algebraic combinatorics, particularly distance-regular graphs, which was also the topic of his PhD thesis. His current research interests also include some other related aspects of algebraic combinatorics, such as Q-polynomial association schemes. He has (co-)authored 5 papers published in international journals, and 2 in the proceedings of international conferences.

In 2015, he started a project, together with Katja Berčič, which would gather various censuses of graphs and make them easily accessible and searchable. They soon realized that there was no reason to limit the collection to just graphs – instead, the project aims to become a repository of various discrete objects. Thus, the project was named *DiscreteZOO*. Janoš Vidali has contributed a Sage package which acts as a framework and interface to the repository, and currently supports accessing graphs from Sage.

He has also developed a package named `sage-drg`, which can be used to compute parameters of association schemes and distance-regular graphs and check for their feasibility. Besides that, he has also contributed to Sage itself, (co-)authoring 8 tickets.

Achievements

[BV18] K. Berčič and J. Vidali. "DiscreteZOO: Towards a Fingerprint Database of Discrete Objects". In: *Mathematical Software – ICMS 2018*. Springer International Publishing, 2018, pp. 36–44. ISBN: 978-3-319-96418-8. URL: https://link.springer.com/chapter/10.1007/978-3-319-96418-8_5.

[EG] *Encyclopedia of Graphs*. URL: http://atlas.gregas.eu (visited on 01/24/2019).

[PSV13] P. Potočnik, P. Spiga, and G. Verret. "Cubic vertex-transitive graphs on up to 1280 vertices". In: *J. Symbolic Comput.* 50 (2013), pp. 465–477. ISSN: 0747-7171. DOI: 10.1016/j.jsc.2012.09.002.

[PSV14] P. Potočnik, P. Spiga, and G. Verret. "On the order of arc-stabilisers in arc-transitive graphs with prescribed local group". In: *Trans. Amer. Math. Soc.* 366.7 (2014), pp. 3729–3745. ISSN: 0002-9947. DOI: 10.1090/S0002-9947-2014-05992-8.

[PSV15] P. Potočnik, P. Spiga, and G. Verret. “A census of 4-valent half-arc-transitive graphs and arc-transitive digraphs of valence two”. In: *Ars Math. Contemp.* 8.1 (2015), pp. 133–148. ISSN: 1855-3966.

Experience

1. During 2008 and 2011, Primož Potočnik led the research project “Catalogue of graphs of high level of symmetry”, funded by the Slovenian Research Agency ARRS. As a part of this project theoretical background was laid for the computer assisted compilation of several exhaustive datasets of graphs and related combinatorial objects. The resulting censuses of highly symmetric graphs by Potočnik and coauthors [PSV13; Con+06; Pot09; PSV15; EET] are now regularly used by a large number of researchers and have already received over 120 citations.
2. Tomaž Pisanski was the Project Leader of the GReGAS project [GG] which was one of the EuroGIGA projects selected by European Science Foundation and financed by national research agencies of participating countries. The main goal of the project was to develop a coherent theory of graph representations with emphasis on symmetric and near symmetric structures or products. One of the activities of the project was an early attempt to make different datasets of combinatorial objects available to the mathematical community in an organised and standardised manner. As a result, the Encyclopedia of Graphs [EG] was developed and implemented as a part of this project. The project started in 2011 and finished in 2014.
3. In 2014 Vladimir Batagelj and his team performed a “network analysis of Zentralblatt MATH data” where the data about works (papers, books) from the time period 1990-2010 that are collected in Zentralblatt MATH database were analysed. The networks were analyzed using Pajek program for analysis and visualization of large networks. The distributions of some properties of works and the collaborations among mathematicians we explored. More details can be found in [CB15].
4. Following the visit of Primož Potočnik of Stephen Wilson of Northern Arizona University, USA in 2003, sponsored by the Fulbright Scholarship, a long term project of compiling a comprehensive on-line census of edge-transitive tetravalent graphs was launched. The project has grown over the years and resulted in a yet another attempt to make the valuable precomputed data available in a user-friendly manner. The current state of the on-line census is available at [EET].
5. More recently, the DiscreteZOO project [BV18] experiments with framework possibilities for datasets of combinatorial objects on the use case of datasets of highly symmetric graphs. It provides two interfaces to the database: a SageMath package [DZb] and a website [DZa].

Infrastructure

The research team at UL has free access to the Slovenian National Supercomuting Grid (SLING). The grid can be used for computationally intensive experiments and large-scale data processing.

4.1.3 EMS: EUROPEAN MATHEMATICAL SOCIETY (FI)

Legal Entity

The European Mathematical Society EMS is a learned society representing mathematicians throughout Europe. It promotes the development of all aspects of mathematics in Europe, in particular mathematical research, relations of mathematics to society, relations to European institutions, and mathematical education, as well ethical issues. EMS is committed to support open availability of mathematical data through its support of the Digital MATH Library DML, its involvement in zbMATH and the EMS Publishing House. The EMS has as its members around 60 national mathematical societies in Europe, 40 mathematical research centres and departments, and 3000 individuals and it is owner of the European Mathematical Society Publishing House.

Within the FAIRMat project EMS will lead, in particular, the activities towards developing a sustainability plan, the dissemination activities, as well as the analysis and establish the legal basis for openly available mathematical data. Prof. Dr. Volker Mehrmann holds a professorship for Mathematics at Technische Universität Berlin and is the current president of EMS (from 2019 to 2022). He will coordinate and supervise the EMS activities within the FAIRMat project.

Curriculum vitae

Volker Mehrmann (PI, male) Volker Mehrmann is full professor for mathematics at TU Berlin. His research interests are in the areas of numerical mathematics/scientific computing, applied and numerical linear algebra, control theory, and the theory and numerical solution of differential-algebraic equations.

He is a member of acatech (the German academy of engineering), academia europaea, president of the European Mathematical Society (EMS), and he was president of GAMM the (International association of Applied Mathematics and Mechanics), chair of MATHEON, the Research Center “Mathematics for key technologies” and chair of the Einstein Center ECMath in Berlin.

He is SIAM Fellow, has received the SIAM Idalia and W.T. Reid Prize, an ERC Advanced Grant and also was member of the ERC Mathematics Panel. He is editor of several journals and editor-in-chief of Linear Algebra and its Applications and coauthor of more than 200 refereed publications and 15 books. He is coordinator of the H2020 Innovative Training Network ’ROMSOC’, Reduced Order Modeling, Simulation and Optimization of Coupled systems.

Achievements

Key publications relevant to the project.

[Ben+99] P. Benner, V. Mehrmann, V. Sima, S. V. Huffel, and A. Varga. “SLICOT-A Subroutine Library in Systems and Control Theory”. In: *Applied and Computational Control, Signals and Circuits* 1 (1999), pp. 499–532.

[Cam+19] S. Campbell, A. Ilchmann, V. Mehrmann, and T. Reis, eds. *Applications of Differential-Algebraic Equations: Examples and Benchmarks*. DAE Forum. Springer Verlag, Heidelberg, 2019.

[CKM12] S. Campbell, P. Kunkel, and V. Mehrmann. “Regularization of linear and nonlinear descriptor systems”. In: *Control and Optimization with Differential-Algebraic Constraints*. Ed. by L. Biegler, S. Campbell, and V. Mehrmann. SIAM, Society of Industrial and Applied Mathematics, 2012, pp. 17–34.

[GLM10] M. Grötschel, K. Lucas, and V. Mehrmann, eds. *Production Factor Mathematics*. acatech and Springer Verlag, 2010.

[KM06] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations — Analysis and Numerical Solution*. EMS Publishing House, Zürich, Switzerland, 2006.

Experience

Volker Mehrmann was PI and chair of the German Science Foundation Excellence Project MATHEON Mathematics for key technologies, which developed application driven basic research in mathematics and its transfer to industry in life sciences, finance, networks, electronic and photonic devices, as well as materials. In the ERC Advanced Grant MODSIMCONMP *Modeling, Simulation and Control of Multi-Physics Systems*, he and his group developed the basic mathematical concepts and algorithms for the modeling, simulation and control of multiscale, multiphysics systems, using the open modeling language Modelica as core. In the nineties he was partner in an EU project which developed the Systems Library in Control SLICOT subroutine library which is sustained still today. He is currently coordinator of the H2020ITN ROMSOC, Reduced Order Modeling, Simulation and Optimization of Coupled systems.

Infrastructure

Not applicable

4.1.4 UPSud: UNIVERSITÉ PARIS-SUD (FR)

Legal Entity

Université Paris-Sud is among the 40 top universities worldwide in the 2013 Shanghai ranking, and is one of the top two French research universities. With about 27000 students, 1800 permanent faculty and 1300 permanent research scientists from national research organisations (CNRS, Inserm, INRA, Inria), it is the largest campus in France. Since 2006, scientists from the University were awarded two Fields medals, one Nobel Prize and a number of other national and international prizes (European Inventor Award 2013, Wolf Prize 2010, Holweck Prize 2009, Japan prize 2007). Université Paris-Sud offers a wide range of qualifications, from the exact sciences to life and health sciences (including medical practice), legal sciences and economics. Research at Université Paris-Sud is an essential part of academic understanding and includes research activities with high commercial potential. Research contracts and partnership with companies make Université Paris-Sud a key actor and a major player in French research. The University is located partly on the Plateau de Saclay, the largest cluster of public and private R&D institutions in France (with ca. 16000 research staff), and is one of the core members of University Paris-Saclay – a world-class university and a world-renowned research and innovation hub.

In the context of this project, Université Paris-Sud is the home of one of the largest group of SAGE developers worldwide. It is a member of the Open Source Thematic Group of the Systematic Paris Region Systems and ICT Cluster.

Curriculum vitae

Nicolas M. Thiéry (leadPI, male) Professor at the Laboratoire de Recherche en Informatique, Nicolas M. Thiéry is a senior researcher in Algebraic Combinatorics with 18 papers published in international journals. Among other things, he is a member of the permanent committee of FPSAC, the main international conference of the domain, a founding member of the upcoming Numfocus Europe non-profit, and a member of Work group on Free and Open Source software for the "Open Science Committee" of the French Ministry for Research. He has collaborators in the US and Canada where he cumulatively spent more than three years (Colorado School of Mines, UC Davis, Providence, Montréal), and in India. He also co-organised fourteen international workshops, in particular SAGE Days, and the semester long program on "Automorphic Forms, Combinatorial Representation Theory and Multiple Dirichlet Series" hosted in Providence (RI, USA) by the Institute for Computational and Experimental Research in Mathematics.

Algebraic combinatorics is a field at the frontier between mathematics and computer science, with heavy needs for computer exploration. Pioneer in community-developed open source software for research in this field, Thiéry founded in 2000 the SAGE-COMBINAT software project (incarnated as MUPAD-COMBINAT until 2008); with 50 researchers in Europe and abroad, this project has grown under his leadership to be one of the largest organised community of Sage developers, gaining a leading position in its field, and making a major impact on one hundred publications[1]. Along the way, he coauthored part of the proposal for NSF SAGE-COMBINAT grant OCI-1147247, and co-organised or taught at a dozen training and dissemination actions (workshops, summer schools, etc.), in America, Africa, Europe, and India.

With 150 tickets (co)authored and as many refereed, Thiéry is himself a core SAGE developer, with contributions including key components of the SAGE infrastructure (e.g. categories), specialised research libraries (e.g. root systems), thematic tutorials, and two chapters of the book "Calcul Mathématique avec SAGE" and its English translation.

Based on this experience, and to tackle the pressing funding needs in the ecosystem of open source mathematical software, Thiéry initiated and lead the European Research Infrastructures project OpenDreamKit #676541 (2015-2019, 15 sites, 50 participants, 8M€), engaging the Kwarc group on board to jointly work on the interaction of Software, Data, and Knowledge.

Software Developer (R, 33 PM) We will hire a full time experienced software developer to work on Tasks **T4.3**, **T3.3**, and **T3.5**, under the leadership of Nicolas M. Thiéry.

The fellow will have a strong software engineering experience, ideally in the Python ecosystem, and some background in databases and mathematics. We further require good communication and team working skills, in particular to work in tight collaboration with international open-source developer communities.

Achievements

1. Leadership of the SAGE-COMBINAT software project since 2000 (formerly MuPAD-Combinat) whose mission is to improve the open source mathematical system Sage as an extensible toolbox for computer exploration in (algebraic) combinatorics, and foster code sharing between researchers in this area.
2. Contribution of more than 500 tickets/100k lines of code to SageMath.

[AST16] A. Ayyer, A. Schilling, and N. M. Thiéry. "Spectral gap for random-to-random shuffling on linear extensions". In: *Experimental Mathematics* (July 2016). arXiv:1412.7488, pp. 1–9. DOI: http://dx.doi.org/10.1080/10586458.2015.1107868.

[1] http://sagemath.org/library-publications-combinat.html, http://sagemath.org/library-publications-mupad.html

[Cas+13] A. Casamayou et al. *Calcul Mathématique avec Sage*. The first of its kind comprehensive introduction to computational mathematics in Sage for education; English translation published in 2018 by SIAM. CreateSpace Independent Publishing Platform, 2013, pp. xii+455. ISBN: 978-1481191043. URL: http://sagebook.gforge.inria.fr/.

[Koh+17] M. Kohlhase, D. Müller, M. Pfeiffer, F. Rabe, N. Thiéry, V. Vasilyev, and T. Wiesing. "Knowledge-Based Interoperability for Mathematical Software Systems". In: *MACIS 2017*. Ed. by J. Blömer, T. Kutsia, and D. Simos. LNCS 10693. Springer Verlag, 2017, pp. 195–210. URL: https://github.com/OpenDreamKit/OpenDreamKit/blob/master/WP6/MACIS17-interop/crc.pdf.

Experience

1. OpenDreamKit (GA No. 676541) Open Digital Research Environment Toolkit for the Advancement of Mathematics, **coordination**.
2. Hosting or coorganisation of dozens of Sage Days (week-long training and development workshops).

Infrastructure

UPSud hosts a local OpenStack based cloud infrastructure CLOUD@VD (400 cores) for its personnel. The participants are regular users of this infrastructure, and in close contact with its maintainers.

UPSud also manages the Digiscope (http://digiscope.fr) network of high-end visualisation platforms and hosts the WILD and WILDER platforms, two ultra-high resolution wall-sized displays with motion capture and touch input for conducting research on collaborative human-computer interaction and visualisation of large datasets.

4.1.5 FIZ: FIZ KARLSRUHE – LEIBNIZ INSTITUTE FOR INFORMATION INFRASTRUCTURE (DE)

Legal Entity

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, founded in 1977, is a non-profit organisation with approx. 300 employees (thereof 40% scientists) and a total budget of about 45 Mio. EUR. FIZ Karlsruhe is a member of the Leibniz Association, one of the major German research organisations comprising 89 research and scientific service institutions. FIZ Karlsruhe has the task to produce scientific-technical information and to provide related services. It also runs internal as well as publicly funded applied research projects. As an international service partner to science and business, FIZ Karlsruhe has strong expertise in handling all issues related to information transfer and knowledge management. FIZ Karlsruhe's business segments complement each other with respect to the information offer and usage possibilities:

- Online service STN International, a leading scientific-technical database service operated in partnership with the American Chemical Society (ACS/CAS);
- e-Research solutions and IT services;
- Databases and information services – producing information systems in broad international cooperations especially in Energy, Mathematics and Crystallography.

Along with the digital transition in science, there is an increasing importance of privacy and intellectual property aspects. In order to develop and provide new information services, we also investigate models and methods for efficient semantic indexing, aggregation, linking, and retrieval of comprehensive heterogeneous and distributed data sources. For both research areas joint professorships have been established together with the Karlsruhe Institute of Technology (KIT).

The following services are in particular related with the proposal:

FIZ Karlsruhe develops and produces various system for mathematical information infrastructure like zbMATH, swMATH. or eLibM. zbMATH (formerly Zentralblatt MATH) is the world's most comprehensive and longest-running abstracting and reviewing service in pure and applied mathematics. It contains about 4 million bibliographic entries with reviews or abstracts since the 19th century, contributed by a community of currently more than 7,000 mathematicians. Currently, zbMATH is in the transition process from the traditional subscription model to an information system providing open services, data and API. The open service swMATH builds upon zbMATH as an information service for mathematical software, systematically linking software packages with relevant mathematical publications. The Electronic Library of Mathematics (eLibM) is one of the largest open digital libraries in mathematics and forms a substantial part of the European Digital Mathematics Library (EuDML).

FIZ Karlsruhe develops and maintains RADAR (Research Data Repository), a multidisciplinary data repository for the archiving and publication of research data resulting from completed scientific studies and projects. Its focus lies on research data of the so-called "long tail", i.e. disciplines that handle smaller amounts of data and usually do not yet have their own infrastructure for research data management. RADAR strives to support researchers and academic institutions in promoting the traceability, reproducibility and transparency of research results. Furthermore, RADAR aims at improving the visibility of research data as independent publications and creates the possibility to reuse data once collected to answer new questions.

FIZ Karlsruhe is the technical operator of the German Digital Library (DDB) and a member of its competence network. It hosts its operating environment and is responsible for data normalization, the loading processes, and the further development of the portal's software.

The group working on *FAIRMat* will be composed of the following non-exhaustive list: Dr. Olaf Teschke, Dr. Moritz Schubotz, Dr. Fabian Müller, and Matthias Razum. Olaf Teschke will serve as principal investigator, the other as researchers and developers.

Curriculum vitae

Olaf Teschke (PI, male, 8 PM) Olaf Teschke is head of mathematics department at FIZ Karlsruhe. This includes responsibility for the world's most comprehensive and longest-running abstracting and reviewing service in mathematics zbMATH; the information service for mathematical software swMATH, and the Electronic Library of Mathematics (eLibM), which also forms an important component of the European Digital Mathematics Library (EuDML). He also lead several projects related to these services on behalf of FIZ Karlsruhe.

He is also Vice Chair of the Committee for Publications and Electronic Dissemination of the European Mathematical Society (EMS), the Global Digital Mathematics Library (GDML) working group of the International Mathematical Union (IMU), and the EuDML Executive Board.

He completed his PhD 2013 in algebraic geometry. He co-organized five international workshops and authored more than 30 refereed publications, mostly in the field of mathematical information and digital mathematics libraries.

Matthias Razum (R/D, male, 4 PM) Matthias Razum is head of e-Research at FIZ Karlsruhe. He is responsible for the development and maintenance of e-Content platforms, virtual research environments, services for research data management and digital preservation, and databases and repository platforms. Notable services operated by his department include the German Digital Library (DDB) and the Research Data Repository RADAR.

He also lead several projects related to these topics and services on behalf of FIZ Karlsruhe, and authored 10 research publications in these areas.

He has served in several national working groups to define research data standards. Currently, he is member of the Advisory Board of the Generic Reseach Data Infrastructure (GeRDI) and member of the Steering Committee of the Preservation and Archival Special Interest Group (PASIG).

Moritz Schubotz (R/D, male, 18 PM) Moritz Schubotz is currently a senior researcher at the Media Technology lab at the University of Wuppertal, Germany. He will join FIZ Karlsruhe in April 2019.

Moritz Schubotz completed his PhD in 2017 in the field of mathematical knowledge management. His achievements include methods for enrichment of formula semantics in digital libraries, enabling significant improvements in findability and interoperability of mathematical objects. In his current research activities, he continues to make mathematics machine readable and pursues several collaborations to make mathematical expressions more useful for humans and computers.

He is an offsite collaborator at NIST (National Institute of Standards and Technology, U.S.A) since 2014 and was a fellow at the National Institute of Informatics in Tokyo, Japan from July 2017 to August 2018. He is a Wikimedia Open Science Fellow and advocates the FAIR principles and Open Science in general. He maintains close connections to the Wikidata community, and he is the leading developer for mathematics support in Wikimedia projects including Wikipedia.

He has published over 30 peer-reviewed articles and contributed to numerous successful grant applications at his former universities.

Achievements

[BRS12] A. Brahaj, M. Razum, and F. Schwichtenberg. “Ontological formalization of scientific experiments based on core scientific metadata model”. English. In: *Theory and Practice of Digital Libraries. Second International Conference, TPDL 2012*. Springer Berlin Heidelberg, 2012, pp. 273–297. DOI: 10.1007/978-3-642-33290-6_29.

[CD18] H. Chrapary and W. Dalitz. “Software products, software versions, archiving of software, and swMATH”. English. In: *Mathematical Software – ICMS 2018. 6th international conference*. Springer, 2018, pp. 123–127. ISBN: 978-3-319-96417-1. DOI: 10.1007/978-3-319-96418-8_15.

[CS17] J. Corneli and M. Schubotz. “math.wikipedia.org: A vision for a collaborative semi-formal, language independent math(s) encyclopedia”. English. In: *AITP 2017. The Second Conference on Artificial Intelligence and Theorem Proving*. 2017, pp. 28–31.

[Kra+16] A. Kraft, M. Razum, J. Potthoff, A. Porzel, T. Engel, F. Lange, K. Van den Broek, and F. Furtado. “The RADAR project – a service for research data archival and publication”. In: *ISPRS International Journal of Geo-Information* 5.3 (2016). ISSN: 2220-9964. DOI: 10.3390/ijgi5030028.

[MT16] F. Müller and O. Teschke. “Intelligent Computer Mathematics”. English. In: *Intelligent Computer Mathematics 2016*. Ed. by M. Kohlhase, M. Johansson, B. Miller, L. de Moura, and F. Tompa. LNAI 9791. Springer, 2016, pp. 63–74. ISBN: 978-3-319-08434-3. DOI: 10.1007/978-3-319-42547-4_5.

Experience

FIZ Karlsruhe has been involved in the creation of research information infrastructure in various projects, and maintains and develops their results sustainably. Connected to the proposal are:

- EuDML, a CIP project to build the European Digital Mathematics Library. Partially funded with EC funds, it built an open collaborative digital library service that collates the distributed content resulting from national digitization programmes.
- swMATH, an information service for mathematical software created in a project funded by the of the Leibniz Association, and further developed within the research campus MODAL funded by the German Federal Ministry of Education and Research.
- RADAR, a project by the German Research Foundation (DFG) to develop a research data infrastructure that facilitates research data management
- MathSearch, a project funded by the Leibniz Association to develop tools for semantic analysis and retrieval of mathematical formula search. In particular, it facilitated the application of the MathWebSearch system to zbMATH and arXiv data.
- The LIMES Project, co-funded by the EC in the Access to Research Infrastructures sector, which transformed Zentralblatt MATH into a distributed European enterprise, and improve and widened its access to European countries by providing structures for a better distribution on the technical and on the economical level.

Infrastructure

FIZ Karlsruhe maintains an ample computational infrastructure to host its services. RADAR is supported by a distributed system in collaboration with the Steinbuch Centre for Computing (SCC) of Karlsruhe Institute of Technology (KIT) and Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) of TU Dresden.

4.1.6 CHA: CHALMERS UNIVERSITY OF TECHNOLOGY (SE)

Legal Entity

Chalmers tekniska hoegskola (Chalmers University of Technology) was founded in 1829 following a donation by William Chalmers, director of the Swedish East India Company, and was transformed into an independent foundation in 1994. Chalmers has developed leading research in the areas of life sciences, materials science, information technology, micro- and nanotechnology, environmental sciences and energy. Chalmers' annual turnover is 3695 million SEK (appr. 383 million EUR), out of which 70 % is related to research. Around 60 % of the research funding is acquired in competition from external sources. Some 13 900 people, including 3258 employees (2955 FTEs), work and study in Chalmers' 13 departments. The university offers PhD and Licentiate programmes as well as MScEng, MArch, BEng and nautical programmes. There are 9502 students (FTE) in programmes leading to 1 335 Master's degrees annually. 1 111 students are involved in doctoral programs leading to about 270 PhD and Licentiate of Technology degrees each year.

Chalmers has an extensive track record of engaging in EU funded research projects. In the Framework Programmes for Research and Innovation, Chalmers has participated in over 750 projects. On a continuous basis Chalmers is involved in approximately 150 EU-projects with an annual EU funding (2017) for research of 22 M€. In Horizon 2020, Chalmers is participating in 131 projects with a contract value of 74 M€ including 9 ERC grants, 30 MSCA grants (15 ITNs) and 26 projects in the Transport challenge. Chalmers is coordinator or single beneficiary of 28 Horizon 2020 projects, including the Graphene Flagship.

The Department of Mathematical Sciences at Chalmers is joint with the University of Gothenburg and has about 200 employees of which about half are senior researchers. The department has strong research groups in complex analysis, number theory, computational mathematics, biomathematics, probability theory and mathematical statistics. The main person involved with *FAIRMat*, Stefan Lemurell, is a member of the number theory group. Relevant for *FAIRMat* is also the fact that the department plays a major role in a Swedish national program in AI spanning over several of the research groups at the department.

Curriculum vitae

Stefan Lemurell (leadPI, male, 12 PM) is an Associate Professor (docent) at the department of Mathematical Sciences at Chalmers and the University of Gothenburg.

He studied mathematics at the University of Gothenburg from 1988 and got his Ph.D. in mathematics in 1997. After the studies he did a postdoc at Rutgers University and at the Institute for Advances Studies in Princeton. From 2000 he has been an Associate Professor at Chalmers.

His research interests have always been in number theory. Currently the main focus is on computational aspects of questions regarding L-functions and the mathematical objects associated to the L-functions.

He is a lead developer and contributor to the L-functions and Modular Forms Database (LMFDB) and has been so since the beginning of the project. Since three years, he is the vice head of the Department of Mathematical Sciences at Chalmers and the University of Gothenburg.

Achievements

[Bob+15] J. Bober, J. B. Conrey, D. W. Farmer, A. Fujii, S. Koutsoliotas, S. Lemurell, M. Rubinstein, and H. Yoshida. "The highest lowest zero of general L-functions". In: *J. Number Theory* 147 (2015), pp. 364–373. ISSN: 0022-314X. DOI: 10.1016/j.jnt.2014.07.023.

[Cre16] J. Cremona. "The L-Functions and Modular Forms Database Project". In: *Foundations of Computational Mathematics* 16.6 (2016), pp. 1541–1553. ISSN: 1615-3383. DOI: 10.1007/s10208-016-9306-z.

[FKL14] D. W. Farmer, S. Koutsoliotas, and S. Lemurell. "Maass forms on $\mathrm{GL}(3)$ and $\mathrm{GL}(4)$". In: *Int. Math. Res. Not. IMRN* 22 (2014), pp. 6276–6301. ISSN: 1073-7928. DOI: 10.1093/imrn/rnt145.

[FKL19] D. W. Farmer, S. Koutsoliotas, and S. Lemurell. "Varieties via their L-functions". In: *Journal of Number Theory* 196 (2019), pp. 364–380. ISSN: 0022-314X. DOI: https://doi.org/10.1016/j.jnt.2018.01.019.

[LM] *The L-functions and Modular Forms Database*. URL: http://www.lmfdb.org (visited on 02/01/2016).

Experience

Stefan Lemurell has had a leading role in the following related projects

- FRG: L-functions and modular forms, NSF grant DMS:0757627 (2008-2012),

- LMF: L-functions and modular forms, EPSRC reference EP/K034383/1 (2013-2019),
- Development of the website www.lmfdb.org,

and has also contributed to SAGE.

Infrastructure

Not applicable

4.1.7 CAE: CAE TECH LIMITED (UK)

Legal Entity

CAE Tech develops bespoke software for engineering applications, combining expertise in simulation, optimisation, geometry and visualisation with the latest cloud software development practises. As a world-leading specialist in Modelica much of the company's work involves the open-standard Modelica language and FMI (Functional Mockup Interface) standard.

Recent software customers and partners include BMR Racing (UK), Claytex Services (UK), Fabber (USA), Onshape (USA), Sanna (Canada), and Theorem Solutions (UK).

A key policy of the company is to support open-source, regularly contributing to open-source software libraries for mathematics or engineering.

CAE Tech also develops Modelica modeling libraries which are distributed to a range of customers in automotive and aerospace industries through partnership with Claytex Services (UK) and Dassault Systemes (FR).

Curriculum vitae

Peter Harman (PI, male, 6 PM) Peter Harman is Founder and Director of CAE Tech. He is a Mechanical Engineer (MEng(Hons) Cardiff University 2001) with significant experience in the engineering simulation software industry.

From 1999 to 2009 Peter applied simulation to engineering problems in the automotive and motorsport industry, working for Lotus Engineering, Ricardo, Romax Technology, Honda Racing F1 and Brawn GP. During this time he initiated the use of the open-standard Modelica modeling language at Ricardo, where it now forms a significant business stream; and led a project to use Modelica across the Honda team.

In 2009 Peter founded Deltatheta UK Limited, a software company that developed a set of tools for modeling and simulation with Modelica, version control of Modelica models, and building of custom tools. Deltatheta provided software and services to customers in automotive and aerospace engineering, including 3 F1 teams and 2 national space agencies.

In 2012 Deltatheta was acquired by CyDesign Labs Inc., a Californian startup. With Peter as Director of Technology, CyDesign delivered to DARPA the first design-space-exploration and system-modeling platform on the cloud, aimed at crowdsourcing the design of military vehicles.

CyDesign was subsequently acquired by multinational simulation software provider ESI Group. Peter was Director of Technology for Virtual Electronics and Systems at ESI from 2013 to 2016, where he led the creation of ESI's first systems-modeling and cosimulation products and their application on automotive customer projects.

Achievements

[Dem+06] M. Dempsey, M. Gäfvert, P. Harman, C. Kral, M. Otter, and P. Treffinger. "Coordinated automotive libraries for vehicle system modelling". In: *Modelica 2006*. 2006. URL: https://www.modelica.org/events/modelica2006/Proceedings/sessions/Session1b2.pdf.

[Har11] P. Harman. "Effective Version Control of Modelica Models". In: 63 (2011), pp. 650–656. ISSN: 1650-3740. URL: http://www.ep.liu.se/ecp/article.asp?issue=063&article=072&volume=#.

[Har13] P. Harman. "Symbolic Application of the Pryce Sigma-Method for Index Reduction of DAEs in CyModelica". In: (2013).

[HT09] P. Harman and M. Tiller. "Building Modelica Tools using the Modelica SDK". In: 43 (2009), pp. 631–636. ISSN: 1650-3740. URL: http://www.ep.liu.se/ecp/article.asp?issue=043&article=072&volume=#.

[TH14] M. Tiller and P. Harman. "recon – Web and network friendly simulation data formats". In: 96 (2014), pp. 1081–1093. ISSN: 1650-3740. URL: http://www.ep.liu.se/ecp/article.asp?issue=096&article=113&volume=#.

Experience

Peter Harman has had significant roles in related collaborative projects:

- Modelica Language Development. Member of the Modelica Association Language Group, working on the specification of the language.
- DARPA AVM Programme (US Department of Defence Funded). Development of cloud-based services for extracting design-space information from Modelica models, and simulating Modelica models.
- "Improving Structural Analysis for Differential-Algebraic Equation Systems" (Leverhulme Trust Funded). Collaboration with Cardiff University and McMaster University, including co-supervision of PhD in Applied Mathematics at Cardiff.

CAE Tech has recently delivered commercial projects related to this work:

- Modelica library static analysis and code coverage tools, for Claytex Services (UK)
- Open-source libraries for accessing CAD data on the cloud, for Onshape (USA)

Infrastructure

CAE Tech uses commercial and open-source simulation tools, and develops software in-house. The company is co-located with other small software companies.

4.2 Third Parties Involved in the Project (including use of third party resources)

Not applicable, as no participant plans to subcontract any tasks.

5 Ethics and Security

5.1 Ethics

We have entered ethical issues in regards to personal data, specifically Question 4 “Personal Data”. Our use of personal data stems from two sources:

- We use publication metadata in our legitimate interest to fulfill the project goal of including publication metadata in our data framework. This data has been made publicly available by the respective authors in order to ensure correct attribution of research data and results, and is expected to be a matter of the public research discourse. While the aggregation and processing of such metadata in the context of the current project proposal has of course not necessarily been foreseen by the data subjects when publishing their data, the potential positive impact on the research community as a whole far exceeds the relatively minor additional loss of control of data that is already processed and aggregated in multiple public and private research infrastructure services. In particular, no fundamental rights or freedoms of the data subjects are violated by this use. All usage of such personal data shall be documented in detail, and information as well as data erasure will be provided upon request in accordance with applicable laws and legislation.
- We use user statistics about our services to the extent that it is necessary to measure the project impact as described in Section 2.

Below we provide a detailed assessment of relevant laws, as well as an itemisation of measures taken to ensure adequate protection of personal data.

Ethics framework and relevant legislation All activities of the *FAIRMat* project will conform to all national and international legislation including

- the Charter of Fundamental Rights of the EU.
- the European Convention for the Protection of Human Rights and Fundamental Freedoms.
- the European Charter for Researchers and the Code of Conduct for the Recruitment of Researchers
- the Data Protection Directive (95/46/EC) of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
- the European General Data Protection Regulation (GDPR)
- the Directive on Privacy and Electronic Communications (2002/58/EC) as well as the new ePrivacy directive.
- the Directive on the Re-use of Public Sector Information (2003/98/EC) as well as the new revised version.

Protection of personal data We plan the following measures to ensure the protection of personal data:

- We will appointment a Data Protection Officer, who will be responsible for overseeing our management of personal data and ensure compliance with legal requirements. This officer will be involved in all project activities, especially where mandated by the GDPR.
- Personal data from publications in datasets designed for research will be anonymized except in cases where the purpose requires full information (e.g., for historical research).
- Where usage statistics or similar are collected, all data will be pseudo-/anonymized, including the replacement of IP addresses and HTTP requests.
- We will prepare a Data Management Plan (see D1.2), which will include all procedures that affect personal data, including (if applicable) the security measures taken to ensure the confidentiality and anonymity of any private personal data.
- We will inform about any data transfers between (EU or non-EU) countries.

5.2 Security

The *FAIRMat* project does NOT involve any of the following:

- activities or results raising security issues: NO
- ‘EU-classified information’ as background or results: NO