Data Ecosystems – Fuelling the Digital Age

Hend*erik* A. Proper[1,2]

[1] Luxembourg Institute of Science and Technology (LIST), Belval, Luxembourg
[2] University of Luxembourg, Luxembourg
E.Proper@acm.org

Abstract. With the increased digitisation of society comes an increase in the role of data. Business analytics, statistics-based AI, the development of digital twins, etc, are typical examples of "data hungry" applications. Such, "data hungry" applications not only need data in different shapes and forms, they also need data from a wide variety of sources.

The systems involved in gathering, storing, processing, analysing, and visualising data, have evolved to be complex systems themselves, involving many actors of widely differing nature. We argue that, as such, these complex systems can be best thought of as 'data ecosystems', which we see as involving the entire complex of social / physical / digital actors which provide, own, sell, buy, exchange, manipulate, store, and use, data. Within these data ecosystems, one needs to deal with technical concerns regarding reliability, performance, interoperability, semantics, etc, as well as social concerns, such as value of data, privacy, trust, ownership, ethics, risk, etc.

In line with this, we argue that there is a need to define / study 'data ecosystems' more closely, where we see a potential future role for the VMBO community.

1 Introduction

Our society is transitioning from the industrial age to the digital age. With the increasing digitisation of society comes an increase in the role of data. Data is gathered from sensors, consequently stored, processed, analysed and visualised, and is eventually consumed by (human and / or digital) actors to enable them to gain insight and / or make informed decisions.

Business analytics, statistics-based AI, the development of digital twins, etc, are typical examples of modern-day "data hungry" applications. For example, data is essential for the training of *statistics-based AI* and the development of *digital twins* [5], while also enabling enterprises to continuously assess their performance in real-time [10] and learn to improve their operations [9]. Industry uses phrases such as *thriving on data* [1] to underline the potential value of data. Meanwhile, we have all grown familiar with the possibilities, as well as the possible positive and negative consequences, of large scale data collection and utilisation as conducted by e.g. Google, Facebook, etc.

The, "data hungry" applications need to be "fuelled" with a wide variety of data resources. For example, ranging from: *raw* observations from different sensors / informants, *processed* and / or *enriched* artefacts in terms of e.g. *predictive* models, representations of *intentions* (e.g. plans, strategy documents, designs, etc), *specifications* (source code, work procedures, etc), or *norms* (regulations, principles, policies, etc). Next to that, such applications also need data from a wide variety of *sources*, requiring the need to transfer ownership of data, or at least a transfer of the right to use the data.

We specifically use the term data, as, in line with e.g. [11, 3], we see information as the *increment* in knowledge / insights which an actor gets when "consuming" data. As such, data are "mere" explicitly represented artefacts that could have *value* to (human and / or digital) actors in the sense that it *may* provide them with relevant / timely *information*.

2 Data ecosystems and their development

As a result of the growing role of data as a key underlying resource, the systems involved in gathering, storing, processing, analysing, and visualising data have evolved to become complex systems themselves, involving different actors with their own interests. We argue that, as such, these complex systems can be best thought of as 'data ecosystems', which we see as involving the entire complex web of social / physical / digital actors (i.e. an ActorWeb [13]), which provide, own, sell, buy, exchange, manipulate, store, and use, data.

Within these data ecosystems, one needs to deal with technical concerns regarding reliability, performance, interoperability, semantics, etc, as well as social concerns, such as value of data, privacy, trust, ownership, ethics, risk, etc. For instance, as the data involved may pertain to (the behaviour of) humans, privacy and ethical considerations may clearly play a role. Furthermore, as the data has some correspondence to "something" in the social, economical, or physical world, it is important to consider quality of this correspondence. At the same time, some actors may have an interest in maliciously changing the data, thus distorting this correspondence. Data also comes with the question of ownership. Data may be of strategic value to some actors, leading them to want to control / sell the access for others. For instance, [4] provides an interesting perspective on this in terms of a personal data market.

A data ecosystem can also be regarded as a "data-management enterprise", i.e. a networked enterprise with "data-management" as its primary business, where data-management refers to all data related activities (gathering, exchanging, manipulating, storing, using, etc.). Such a "data-management enterprise" will typically be embedded in a larger enterprise, where the latter focuses on a "regular" products / services.

The development of data ecosystems, as "data-management enterprises", can clearly benefit from the use of enterprise modelling approaches. As such, the above considerations directly apply, while at the same time suggesting the need to more specifically capture data ownership, data lineage, value of data (to specific stakeholders), access control, data regulations, etc.

3 Research challenges

We conclude this discussion paper with some some possible research challenges in relation to data ecosystems. They are certainly not intended as a complete list of challenges, but should rather provide a starting point for a broader discussion at the VMBO workshop.

Data as a key resource – It is clear that data is a key resource in a data ecosystem. As such, it generates several important questions:
What is the (potential) value of data? How to assess / express this?
What does ownership of data mean, also in relationship to "the original" (e.g. behaviour / properties of a human being), and associated privacy concerns.
How to model the ownership, access to, the (potential!) value of data, etc, as well as associated risks?
How to take these elements into due consideration when designing / developing / evolving data ecosystems?

Trust at the core – Exchanging data requires trust between the (human) actors involved, regarding (1) the way they handle the data and / or access to the data, (2) on how the data is gathered (quality of data), and (3) the way data is used (ethics and privacy). This results in several challenges:
What is "trust" in the context of data ecosystems, and what can threaten such trust?
How to conduct a risk analysis on how data is handled?
How to nurture / increase trust between different stakeholders?
Does the notion of "privacy by design" work in an (open and evolving) data ecosystem?
How to identify system risks for data ecosystem, and how to manage these?

Regulation of data ecosystems – Regulators are likely to have a need to regulate the risks (see above), privacy concerns of data ecoststems, as well as possibly other properties. This results in challenges such as:
To what extent can data ecosystems be regulated at all, given their open, and evolving, nature?
How to express, and enforce, regulations on data ecosystems?
What are the possible risks that need regulation?

Data needs semantics – With the large amounts of data available to us, it is important to also capture its informational semantics. Both to enable re-use and relating (interoperability between) different data sources. Of course this takes us back to *semantic modelling* [8] and *information modelling* [2, 12], as well as (foundational) ontology approaches [6, 7]. This leads to the following broad challenge:
How to re-apply old *(but proven) semantic / information / ontology modelling approaches to continuously capture the semantics of (evolving) data streams flowing between the web of actors involved in a data ecosystem?*

From data to information – Data, in itself, is "just" a passive resource. Even enriched data (e.g. predictive models, digital twins, etc) is. Data does not become "activated" until an actor (human or digital) becomes *informed* by it in the context of learning, decision making, etc. In doing so, the actor "gleans" information from the data (as a potential information carrier [14]). In the context of "the web", finding the right data carriers to relinquish one's information need was already a major challenge. In the context of data ecosystems, this challenge will only grow, leading to the following broad challenges: *How to evolve / extend existing search / discovery techniques form information retrieval / discovery towards data ecosystems?*

How to apply different techniques for visualisation, verbalisation, audiofication, etc, to make data better accessible to human actors, to increase the information they may glean from the data?

References

1. Capgemini. TechnoVision 2012 – Bringing Business Technology to Life. Research report, Utrecht, the Netherlands, 2009.
2. P. P. Chen. The Entity–Relationship Model: Towards a Unified View of Data. *ACM Transactions on Database Systems*, 1(1):9–36, March 1976.
3. E. D. Falkenberg, A. A. Verrijn–Stuart, K. Voss, W. Hesse, P. Lindgreen, B. E. Nilsson, J. L. H. Oei, C. Rolland, and R. K. Stamper, editors. *A Framework of Information Systems Concepts*. IFIP WG 8.1 Task Group FRISCO, IFIP, Laxenburg, Austria, 1998. ISBN: 3-901-88201-4
4. R. Farrelly and E. K. Chew. Designing a primary personal information market as an industry platform: a service innovation approach. In *Hawaii International Conference on System Sciences 2017 (HICSS)*, 01 2017.
   `doi:10.24251/HICSS.2017.556`
5. M. Grieves. Virtually Intelligent Product Systems: Digital and Physical Twins. In S. Flumerfelt, K. G. Schwartz, D. Mavris, and S. Briceno, editors, *Complex Systems Engineering: Theory and Practice*, pages 175–200. American Institute of Aeronautics and Astronautics, 2019. ISBN: 978-1624105647
6. N. Guarino. Formal Ontology and Information Systems. In N. Guarino, editor, *Proceedings of FOIS'98, Trento, Italy*, pages 3–15, Amsterdam, the Netherlands, June 1998. IOS Press.
7. G. Guizzardi. On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models. In O. Vasilecas, J. Eder, and A. Caplinskas, editors, *Databases and Information Systems IV - Selected Papers from the Seventh International Baltic Conference, DB&IS 2006, July 3-6, 2006, Vilnius, Lithuania*, volume 155 of *Frontiers in Artificial Intelligence and Applications*, pages 18–39. IOS Press, 2006. ISBN: 978-1-58603-715-4
8. M. Hammer and D. McLeod. Database Description with SDM: A Semantic Database Model. *ACM Transactions on Database Systems*, 6(3):351–386, September 1981.
9. E. D. Hess. *Learn or Die: Using Science to Build a Leading-Edge Learning Organization*. Columbia University Press, 2014. ISBN: 978-0231170246
10. M. H. Hugos. *Building the Real-Time Enterprise: An Executive Briefing*. Wiley, Hoboken, New Jersey, 2004. ISBN: 978-0471678298
11. B. Langefors. *Editorial notes to: Computer Aided Information Systems Analysis and Design*. Studentlitteratur, Lund, Sweden, 1971.
12. G. M. Nijssen and T. A. Halpin. *Conceptual Schema and Relational Database Design: a fact oriented approach*. Prentice Hall, Englewood Cliffs, New Jersey, 1989. ISBN: 0-13-167263-0
13. H. A. Proper. Fundamentally understanding IT? - Why Web 2.0 needs architects. Part II, 2008.
    `http://tinyurl.com/mc3ozv8`
14. H. A. Proper and P. D. Bruza. What is information discovery about? *Journal of the American Society for Information Science*, 50(9):737–750, July 1999.
    `doi:10.1002/(SICI)1097-4571(1999)50:9<737::AID-ASI2>3.0.CO;2-C`