

SAKE: A Semantic Authoring and Annotation Tool for Knowledge Extraction

Jan Grau¹, Kimberly Garcia¹ and Simon Mayer¹

¹*Institute of Computer Science, University of St.Gallen, Switzerland*

Abstract

Greenhouse Gas (GHG) accounting is traditionally a lengthy and manual process that requires the expertise of experienced environmental scientists; due to the recognition of the climate crisis through upcoming regulations on GHG accounting around the planet, the demand for tools that can support these environmental experts and accelerate their work is growing considerably at the moment. GHG accounting is merely one application of automated support tools that require the preservation of expert knowledge in a machine-readable and machine-understandable format; across fields, this is highly relevant for automating processes that today can only be performed by individuals with specialized training. In this paper, we present SAKE, a Semantic Authoring and Annotation tool for Knowledge Extraction that allows domain experts with no proficiency in semantic technologies annotating domain-specific PDF files, creating a Knowledge Graph with instances of standardized (or new) ontologies. The resulting Knowledge Graph can then be integrated into systems to automate specialized processes. SAKE has been developed together with domain experts in the field of environmental science and is currently used in the scope of a joint project on GHG accounting.

Keywords

Semantic Authoring, Semantic Annotator, PDF annotator, Semantic Web Tool.


1. Introduction

From science to law and from research papers to regulatory documents, a large amount of textual knowledge is today available in the form of PDF files. The knowledge transported through these PDFs, while valuable for appropriately contextualized human readers, today remains hard to integrate with automated systems. While current machine-learning methods, such as large language models, mitigate this problem for content that aligns well with their training data, these fall short for specialized knowledge that requires contextualized processing. Such contextualization could be achieved if the information in a PDF was semantically integrated with shared ontologies. This would not only enable automatic processing of the content, but also—in-line with the core tenet of the Semantic Web—support the interlinking of pieces of information across documents, institutions, and domains. While semantic annotation is readily supported for HTML content, e.g., with Web-Annotation-based tools such as dokieli [1], PDF documents today remain sidelined in Semantic Web tooling. There are good historical, technical, and social reasons for this; however, given the wide range of domains and large amount of

SEMANTICS'24: Posters and Demos, September 17–19, 2024, Amsterdam, Netherlands.

✉ janerik.grau@student.unisg.ch (J. Grau); kimberly.garcia@unisg.ch (K. Garcia); simon.mayer@unisg.ch (S. Mayer)

ORCID [0009-0006-0565-2034](https://orcid.org/0009-0006-0565-2034) (J. Grau); [0000-0002-4971-2944](https://orcid.org/0000-0002-4971-2944) (K. Garcia); [0000-0001-6367-3454](https://orcid.org/0000-0001-6367-3454) (S. Mayer)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

information available (often exclusively) through PDFs, we argue that it is time to pull PDF-based communities into the world of Knowledge Graphs. Thus, we created SAKE, a **Semantic Authoring and Annotation tool for Knowledge Extraction** that permits semantically lifting PDF documents through ontology-based annotations generated by a user, thereby simplifying the integration of information in PDF documents into the Semantic Web. The development of SAKE was motivated by an innovation project that aims at automating GHG accounting through Semantic Web technologies¹. In this contribution, we introduce SAKE's implementation and features, and we discuss the GHG accounting project that is currently taking advantage of SAKE.

2. The SAKE Annotation Tool

SAKE is a Web application built upon PDF.js², a library developed by Mozilla that provides all the functionalities of a PDF reader. SAKE defines its own skin to offer semantic annotation functionalities (see Figure 1(c)) next to the full capabilities of PDF.js. Specifically, SAKE enhances the PDF highlighting functionality to allow users to transform relevant content found in a document into structured knowledge expressed in the Resource Description Framework³ (RDF). SAKE's current implementation uses AtomicData⁴ as a semantic back end. AtomicData hosts ontologies used for annotating documents and user data to add provenance information to annotations. In our implementation, AtomicData could be easily replaced by any other user-based graph database, such as Solid⁵ or GraphDB⁶.

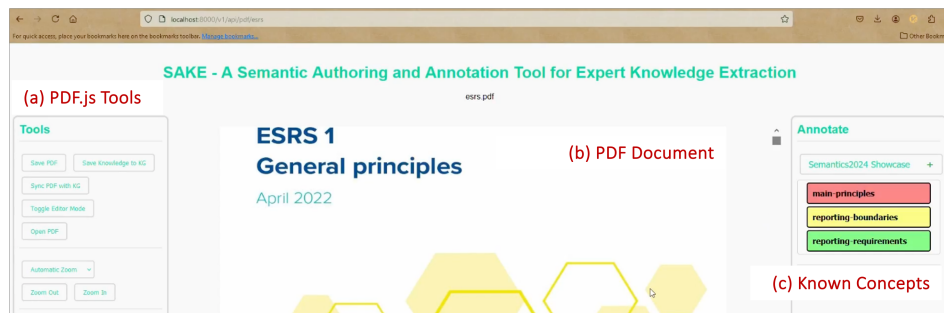


Figure 1: SAKE. (a) functionalities associated with a PDF reader (e.g., next page and zoom); (b) PDF to annotate; and (c) user *Known Concepts* from an ontology loaded through the semantic backend.

To annotate a PDF file, a user (we consider domain experts) first loads an ontology (expressed in RDF) into SAKE's semantic back end. The classes specified in this ontology are considered the user's *Known Concepts* (KCs). SAKE displays all KCs on the right side of the user interface (see Figure 1). To annotate a PDF entity (text or figure), the user selects a KC and then selects the PDF entity. Then, SAKE displays a pop-up window that prompts the user for additional information

¹<https://wiser-climate.com/>

²<https://mozilla.github.io/pdf.js/>

³<https://www.w3.org/RDF/>

⁴<https://atomicdata.dev/>

⁵<https://solidproject.org/>

⁶<https://graphdb.ontotext.com/>

(see Figure 2) corresponding to the attributes and relationships related to the selected KC (i.e., object and data properties) and specified in the loaded ontology.

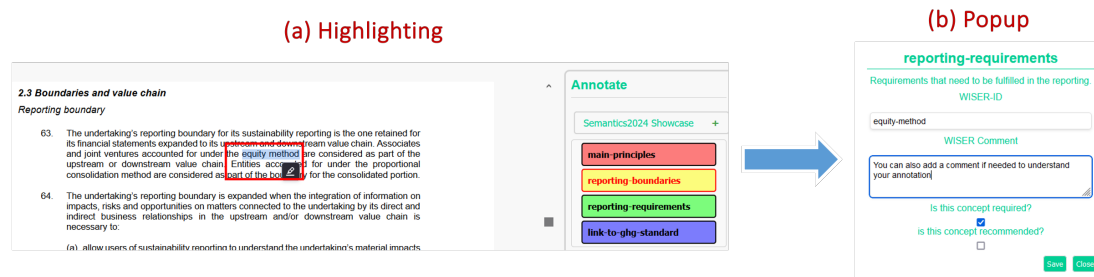


Figure 2: (a) Highlighting tool after selecting a *Known Concept*; (b) Popup window to add more annotation information.

To ensure compatibility with all common PDF readers (e.g., Adobe Acrobat), the annotation is stored as an RDFa string in the PDF document’s *Content* dictionary (cf. the PDF specification [2]). Hence, the PDF document can be distributed with the embedded structured data, and collaborators not using SAKE will still see the semantic annotations when using other PDF readers. While the semantic annotations (being RDF) may be hard to read, they can still be modified with any common PDF reader. Since SAKE embeds semantic annotations within a PDF file, it acts as a self-contained Knowledge Graph. Thus, SAKE RDFa annotations can immediately be used with Semantic Web applications, such as dokieli [1]. Moreover, when an expert shares an annotated PDF file with a colleague using SAKE, this colleague is able to read the text that surrounds an annotation, providing them with context and improving their understanding of a KG that has been created in a collaborative fashion.

Furthermore, SAKE integrates a Web server and responds to HTTP requests that specify appropriate content types (e.g., *text/turtle* or *application/ld+json*) with the graph embedded within the currently open PDF document. Finally, SAKE provides domain experts with the means to add new concepts and properties to existing ontologies; when a new concept is added, SAKE asks the expert to specify the corresponding HTML elements (e.g., a text field or a drop-down menu) to be displayed in the annotation pop-up window that is associated with the concept. This pop-up information is stored as a list of RDF instructions, which SAKE interprets at runtime. These instructions include validating and mapping strings to concepts in the ontology.

3. SAKE for GHG Accounting

Today, GHG accounting is a time-consuming and expensive process that requires highly specialized environmental scientists to manually analyze companies’ processes, including their supply chains; even large multinational companies only commission these assessments rarely due to the amount of manual and expensive effort required. Faster and more cost-effective GHG assessment is required not only to comply with sustainability reporting obligations (e.g., the Swiss Ordinance on Climate Disclosures), but also to regularly assess current practices and reconsider company strategies to reach decarbonizations goals. In this context, WISER is

an interdisciplinary project⁷ coordinated by Empa (Swiss Federal Laboratories for Materials Science and Technology) that aims at providing technological tools to increase the efficiency of GHG assessments. The project specifically required a way to capture knowledge from PDF documents as contextualized by the environmental scientists at Empa in a machine-understandable way. This applies primarily to *Assessment Standards* documents that must be followed when creating a GHG assessment and are published by different organizations (e.g., ISO, the European Commission, the World Business Council for Sustainable Development, or the World Resources Institute), which use idiosyncratic nomenclature and inconsistent concept definitions. Hence, two GHG assessment reports might not be comparable if different standards were followed or even if the same was followed but interpreted differently.

To increase reproducibility and consistency across GHG assessment reports, WISER aims to create ontologies that describe different assessment standards and bridge ontologies that identify commonalities that permit the automatic translation of reports across assessment standards. Given that the environmental experts in our team are not ontologists, SAKE is proving value in capturing their knowledge when reading an assessment standard. The KG resulting from experts annotation will be incorporated in a Web application that accelerates the creation of GHG assessments and can *translate* reports from one assessment standard to another.

4. Related Work

Providing non-semantic technologies experts with tools for creating semantically enriched content has remained a challenge for several decades [3]. Early tools focused on bringing the Semantic Web vision forward by, for example, annotating Web content with metadata. Such is the case of Annotea [4], which provided infrastructure to make remarks (in RDF) on content available on the Web, at the resource level, or on selected text (e.g., add the place in which a picture was taken). Loomp [5] was a system for serving RDF or XHTML content, it proposed the One Click Annotator, that allowed specialist (e.g., journalists) creating semantically enriched documents (e.g., news articles), linking them to data sources, and sharing them with other colleagues for further annotation or for publishing. Semantator [6] is a Protégé plugin for annotating biomedical data that provides semi-automatic annotation support using domain ontologies. SlideWiki [7] provides manual and semi-automatic annotation tools for enriching slide decks with linked data. It allows adding slide deck metadata or linking the content of a slide to DBpedia entries. Dokieli [1] is a platform for decentralized authoring, annotating, and publishing HTML documents while engaging in social interactions. Dokieli uses HTML+RDFa to edit documents and discuss them collaboratively. Sangrahaka [8] is a Web application that allows administrators to create a schema used by annotators; curators can then verify annotations and resolve conflicts. Similarly, SenTag [9] is a Web application that allows users creating XML annotations on plain text.

As described, most of the relevant related tools focus on HTML content, not on PDF documents as SAKE does). These documents hold vast amount of knowledge if read and annotated by experts. Moreover, SAKE is highly interested in high quality semantic annotations to integrate them in a tool (e.g., a dashboard) that can accelerate highly specialized real-world processes.

⁷<https://wiser-climate.com/>

5. Conclusions and Future Work

To overcome one relevant entry barrier to using semantic technologies by domain experts, we have created SAKE, a tool that allows domain experts to create structured knowledge from PDF documents. This knowledge can then be exported as a KG and integrated into a tool for supporting highly specialized tasks such as GHG accounting. SAKE is provided with this publication as open source⁸, and remains in iterative development; it is currently used by environmental scientists in the scope of an interdisciplinary GHG accounting project. However, we expect the need for semantic annotation, sharing, and automated reasoning on top of extracted knowledge to keep growing across a variety of domains in which knowledge is still documented in PDFs, and their interpretations remain within the experts' minds.

Acknowledgments: We thank Dr. Didier Beloin-Saint-Pierre, Alexander Kirsten, and Dr. Daniel Lachat, environmental scientists at Empa, for their support in testing SAKE. SAKE has been developed as part of the WISER flagship project funded by Innosuisse.

References

- [1] S. Capadisli, A. Guy, R. Verborgh, C. Lange, S. Auer, T. Berners-Lee, Decentralised authoring, annotations and notifications for a read-write web with dokieli, in: *Web Engineering*, Springer International Publishing, 2017. doi:https://doi.org/10.1007/978-3-319-60131-1_33.
- [2] Document management – Portable document format – Part 2: PDF 2.0, 2020. URL: <https://www.iso.org/standard/75839.html>.
- [3] S. Handschuh, S. Staab, F. Ciravegna, S-CREAM – Semi-automatic CREAtion of Metadata, in: *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, Springer Berlin Heidelberg, 2002. doi:[10.1007/3-540-45810-7_32](https://doi.org/10.1007/3-540-45810-7_32).
- [4] J. Kahan, M.-R. Koivunen, Annotea: an open RDF infrastructure for shared Web annotations, in: *Proceedings of the 10th international conference on World Wide Web*, ACM, Hong Kong Hong Kong, 2001. doi:[10.1145/371920.372166](https://doi.org/10.1145/371920.372166).
- [5] M. Luczak-Rosch, R. Heese, Linked data authoring for non-expert, in: *Linked Data on the Web Workshop*, 2009. URL: https://ceur-ws.org/Vol-538/ldow2009_paper4.pdf.
- [6] C. Tao, D. Song, D. Sharma, C. G. Chute, Semantator: Semantic annotator for converting biomedical text to linked data, *Journal of Biomedical Informatics* 46 (2013). doi:[10.1016/j.jbi.2013.07.003](https://doi.org/10.1016/j.jbi.2013.07.003).
- [7] A. Khalili, K. A. de Graaf, SlideWiki – A Platform for Authoring FAIR Educational Content, in: *SEMANTiCS (Posters & Demos)*, 2018.
- [8] H. Terdalkar, A. Bhattacharya, Sangrahaka: a tool for annotating and querying knowledge graphs, *ACM*, 2021. doi:[10.1145/3468264.3473113](https://doi.org/10.1145/3468264.3473113).
- [9] A. Loreggia, S. Mosco, A. Zerbinati, SenTag: A Web-Based Tool for Semantic Annotation of Textual Documents, *Proceedings of the AAAI Conference on Artificial Intelligence* (2022). doi:[10.1609/aaai.v36i11.21724](https://doi.org/10.1609/aaai.v36i11.21724).

⁸https://github.com/jangrau13/semantics2024_sake