# Towards research infrastructures that curate scientific information: A use case in life sciences

Markus Stocker[1,2][0000−0001−5492−3212], Manuel Prinz[1][0000−0003−2151−4556],
Fatemeh Rostami[3][0000−0002−0992−6227], and Tibor Kempf[3]

[1] TIB Leibniz Information Centre for Science and Technology
Welfengarten 1 B, 30167 Hannover, Germany
{markus.stocker,manuel.prinz}@tib.eu
[2] MARUM Center for Marine Environmental Sciences
PANGAEA Data Publisher for Earth & Environmental Science
Leobener Strasse 8, 28359 Bremen, Germany
mstocker@marum.de
[3] Division of Molecular and Translational Cardiology
Department of Cardiology and Angiology, Hannover Medical School
Carl-Neuberg-Strasse 1, 30625 Hannover, Germany
{rostami.fatemeh,kempf.tibor}@mh-hannover.de

**Abstract.** Scientific information communicated in scholarly literature remains largely inaccessible to machines. The global scientific knowledge base is little more than a collection of (digital) documents. The main reason is in the fact that the document is the principal form of communication and—since underlying data, software and other materials mostly remain unpublished—the fact that the scholarly article is, essentially, the only form used to communicate scientific information. Based on a use case in life sciences, we argue that virtual research environments and semantic technologies are transforming the capability of research infrastructures to systematically acquire and curate machine readable scientific information communicated in scholarly literature.

**Keywords:** Scientific information · Scholarly communication · Knowledge representation · Virtual research environments · Research infrastructures · Knowledge infrastructures

## 1 Introduction

The critique is not new and the quest remains: Despite advances in information technology and systems, the format of the scholarly article has largely remained unchanged [16, 17, 32]. The wealth of scientific information conveyed by the steadily increasing number of published articles [43, 9, 27] continues to be confined to the document, seemingly inseparable from the medium as hieroglyphs carved in stone.

Document centric scholarly communication has its challenges. Most obviously, machine processing of the information communicated in scholarly articles is very limited. While words can be indexed and searched, the semantics of numbers, text, figures, symbols, etc. are hardly accessible to computers and modern exploration, retrieval, question answering and visualization thus not applicable. Such limited machine support hinders the efficient processing of literature since relevant information is “buried” in documents and finding information relies on sifting through documents. Given the growing scientific output, processing literature ties up increasing resources.

To be sure, important advances have been made. The interlinking of articles with related entities is a notable recent development. Aided by interoperable information infrastructures—such as DataCite, Crossref, literature and data publishers—articles are increasingly linked to related persistently identified datasets, audio/video, samples, instruments, software, people, institutions. The Scholix framework for scholarly link exchange [10] is a project that focuses on interoperability of information about the links between scholarly literature and data. Related advancements can be noticed also in systems that are well-known to researchers. Taking the link between articles and citations as an example, ResearchGate now shows citations “in context” by pointing readers directly to the relevant position in articles. Other related projects include Research Graph [3], RMap [25], and Research Objects [8]. The resulting graphs enable new forms of information publication, search, navigation and discovery. However, it is not scientific information communicated in scholarly literature that these graphs capture but information (i.e., metadata) about the digital objects used in communication and their relationships to contextual entities.

Another notable development is in technologies and vocabularies for machine readable representations of scientific information authors communicate in scholarly literature. Indeed, representing scientific knowledge claims has been explored for at least a decade. With the HypER approach, de Waard et al. [18] proposed to extract knowledge from articles “to allow the construction of a system where a specific scientific claim is connected, through trails of meaningful relationships, to experimental evidence.” García-Castro et al. [22] proposed an extension to the Annotation Ontology [15] that enables the modelling of concepts and relations of scholarly articles, such as ‘claim’, ‘hypothesis‘ or ‘contradicts’ and ‘proves’. Nanopublications [31, 23] is a concept and model designed to represent, in machine readable form, scientific statements. The OBO Foundry [34] publishes ontologies that include numerous relevant concepts e.g., for the machine readable representation of statistical hypothesis tests or average values. As a result, it is now possible to describe scientific information authors communicate in scholarly literature in machine readable form and thus have infrastructures curate, process, and publish such information as distinct information objects.

A third important advancement is in virtual research environments (VREs) [2, 12] (also known as virtual laboratories and science gateways) that enable the execution of data analysis on interoperable infrastructure. Since VREs can be extended in functionality and engineered to meet advanced requirements, the p-value resulting in a statistical hypothesis test is no longer a mere number (as is generally the case in local computational environments) but can be an information object relating the p-value to the kind of statistical test performed, the involved continuous variables and values, and even data provenance in laboratory experiments. In other words a machine readable description of the performed statistical hypothesis test.

Based on a use case in life sciences, we argue that key technologies needed for research infrastructures to acquire and curate more of the scientific information communicated in scholarly literature as machine readable interlinked yet distinct information objects are in place. While certainly challenging, technological integration seems to be on the horizon. Here, we depict such an integration in the context of an open project[4] recently initiated by the TIB Leibniz Information Centre for Science and Technology which aims to develop infrastructure that acquires, curates, and processes scientific information communicated in scholarly literature [5]. In addition to technical considerations elucidated on the use case, we discuss possible pathways through which machine readable scientific information may be systematically acquired by the prospective infrastructure. We also present recent developments and some near-future plans of the project.

## 2 Use Case

We aim to reproduce and represent, in machine readable form, the statistical hypothesis test supporting the scientific statement that "IRE binding activity was significantly reduced in failing hearts" as published by Haddad et al. [24, p. 364] in their article entitled *Iron-regulatory proteins secure iron availability in cardiomyocytes to prevent heart failure* recently published by European Heart Journal.

Iron-responsive elements (IREs) are conserved nucleotide sequences located in uncoded regions of iron-related transcripts (mRNA). These elements can be bound by iron-regulatory proteins (IRPs) in order to regulate the iron homeostasis in cells, which is essential for cell survival since iron is a key co-factor for many enzymes involved in numerous biological processes, ranging from DNA synthesis to energy metabolism. In iron-depleted cells, IRP activity increases in order to secure the iron availability [26]. According to Haddad et al., patients with heart failure (a condition whereby the heart is unable to pump sufficiently) show reduced IRP activity and iron content in heart cells, leading to impaired heart function.

The statement by Haddad et al. is based on data reported in the plot shown in Figure 1B, specifically for non-failing hearts (NF) and patients with failing heart (F). The data reported in the plot are themselves sourced in the electrophoretic mobility shift assay shown as image in Figure 1B. The quantification of the image is done using ImageJ [33], an image processing and analysis software.

Given the data, Haddad et al. use Prism (GraphPad Software) to perform a Student's t-test and find the reported statistical difference ($P < 0.001$) in mean IRE binding activity between the two groups (NF and F). Hence the author's

[4] Open Research Knowledge Graph: `http://orkg.org` (Accessed: October 16, 2018)

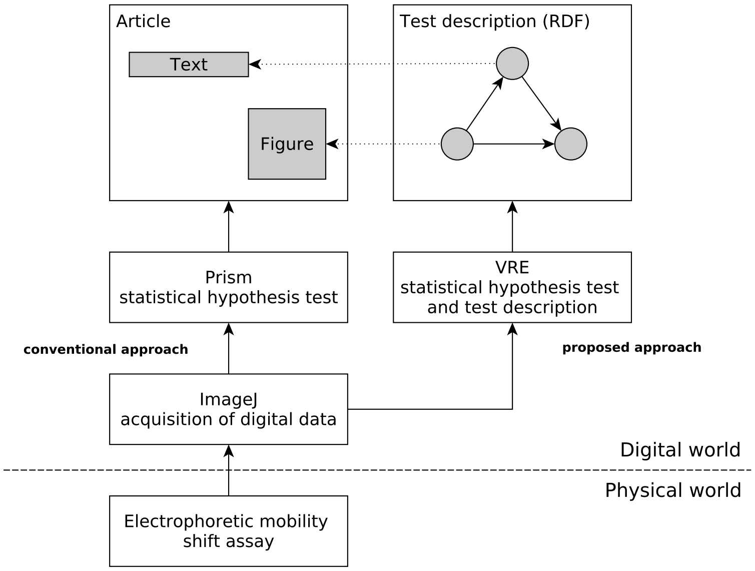

**Fig. 1.** Overview of the main aspects of the conventional and proposed approaches.

statement that "IRE binding activity was significantly reduced in failing hearts." Prism is also used to create the plot shown in Figure 1B.

## 3 Architecture

Figure 1 contrasts the main aspects of the conventional approach just described with those of the proposed one. Akin to the conventional approach, the proposed one adopts a system architecture with technical and social subsystems, and sociotechnical subsystem integration. However, subsystems differ in details.

In the proposed approach, the technical subsystem consists of a digital infrastructure that operates a semantically enhanced Virtual Research Environment (VRE). While VREs typically support numerous features e.g., cataloguing and communication, of primary concern here is a component for data analysis. It is this VRE component that we suggest to semantically enhance. The technical subsystem also consists of a component capable of storing and retrieving information objects. The social subsystem consists of individual researchers, members of research communities. Among other activities, researchers are the agents that perform data analysis. The proposed approach also relies on sociotechnical integration. Indeed, researchers are required to move data analysis from local computing environments into the VRE. This is to ensure that the data derived in analysis conform with the representational requirements of the system.

Data analysis is the key activity that evolves uninterpreted data to scientific information, ultimately published in scholarly literature. We borrow the notion

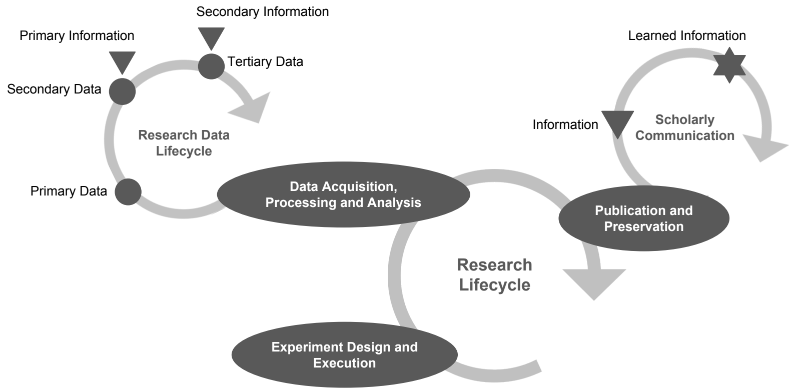

**Fig. 2.** From uninterpreted data to scientific information in the research lifecycle.

of *data interpretation* from the unified definitional model of data, information, and knowledge proposed by Aamodt and Nygård [1]. According to the model, data are uninterpreted symbols with "no meaning for the system concerned" and are input to an interpretation process. Information is interpreted data i.e., data with meaning and the output from data interpretation. Interpretation occurs "within a real-world context and for a particular purpose." Aamodt and Nygård's model also defines knowledge as learned information. As the output of learning processes, "knowledge is information incorporated in an agent's reasoning resources."

Floridi [21] further elaborates the definition of information. Building on a widely adopted General Definition of Information (GDI), he develops a definition of semantic information. GDI defines information in terms of "data + meaning." Floridi proposes a more precise formulation that borrows the term *infon* [7, 19], a discrete item of information. The infon $\sigma$ is an instance of information, understood as *semantic content*, if and only if $\sigma$ consists of $n$ data, $n \geq 1$; the data are well formed; and the well-formed data are meaningful (i.e., of significance to some person, situation or machine). Of specific interest here is *factual* semantic content i.e., semantic content about a situation or fact that can be qualified as either true or false. Only semantic content that is true is informative. Thus, Floridi suggests that $p$ qualifies as factual *semantic information* if and only if $p$ is well-formed, meaningful, and *truthful* data. Furthermore, Floridi proposes a classification of types of data, of which two are of importance here. *Primary data* are the principal data stored, for example in a database while *derivative data* are data that "can be extracted from some data whenever the latter are used as indirect sources in search of patterns, clues or inferential evidence about things other than those directly addressed by the data themselves."

**Listing 1.** Python implementation of the statistical hypothesis test.

```
import numpy as np
import pandas as pd
from scipy.stats import ttest_ind

labels = ['non-failing heart (NF)', 'failing heart (F)']
data = [(99, 52), (96, 40), (100, 38), (105, 18),
        (np.nan, 11), (np.nan, 5), (np.nan, 42),
        (np.nan, 55), (np.nan, 53), (np.nan, 39),
        (np.nan, 42), (np.nan, 50)]

df = pd.DataFrame.from_records(data, columns=labels)
tt = ttest_ind(df['non-failing heart (NF)'],
               df['failing heart (F)'],
               equal_var=False, nan_policy='omit')

tt.pvalue
```

Figure 2 places these concepts in the context of the research lifecycle. Uninterpreted, primary data resulting in observation, experimentation, or simulation activities enter the research data lifecycle by data acquisition. Primary data may be processed in activities other than data interpretation (e.g., aggregation or interpolation). In such activities, derived data remain uninterpreted and without meaning for the system concerned. It is in data analysis that data are interpreted and derived data are information, meaningful and—following Floridi—truthful data for the system concerned. Along research data lifecycles, data may be processed and analysed repeatedly resulting in secondary, tertiary, quaternary, etc. data and, if data are meaningful and truthful for the system concerned primary, secondary, etc. information.

Factual semantic information is a fundamental unit in scholarly communication. Figure 2 suggests that information is learned, incorporated in an agent's (researcher, primarily) "reasoning resources" (knowledge base). Through learning processes, in scholarly communication information thus evolves to knowledge, specifically learned scientific or scholarly information.

Instances of factual semantic information and learned scientific information communicated in scholarly literature are the objects which the proposed architecture aims to represent, acquire, curate, and publish for further reuse. Their representation is machine readable. Critically, not just the data that constitute information are machine readable: Meaning is machine readable, too. Hence, not only is the value `0.013` machine readable but so is its meaning as a e.g., p-value. We now present the implementation of the use case following this architecture and conceptual framework.

## 4 Implementation

We implement the statistical hypothesis test using Jupyter [29] in Python, specifically Jupyter Lab, the next-generation web-based interface for Project Jupyter. Jupyter Lab acts as VRE component that provides services for data analysis among the range of services typically provided by a full-fledged VRE e.g., D4Science VREs [11].

The complete Jupyter notebook is published [38]. We limit the presentation here to the key elements. Given the experimental data, the statistical hypothesis test can be easily implemented using SciPy [28]. Listing 1 shows the implementation in detail. The last line returns the computed p-value i.e., 0.0000000131. This merely reproduces in Jupyter Lab some of the output researchers obtain using Prism.

More interesting is the possibility to describe, in machine readable form, the performed statistical hypothesis test. Since our Jupyter Lab based VRE component can be extended with novel functionality, we implement a function that returns a description of the test in RDF (Resource Description Framework) [30]. Listing 2 displays the core of the description (prefixes are omitted). The numeric p-value is described as the output of a two sample t-test with unequal variance (`STATO_0000304`). The test description also specifies iron-responsive element binding (`GO_0030350`) as the study design dependent variable (`OBI_0000751`), a specified input of the statistical hypothesis test. Omitted here for the sake of brevity, the description also includes the continuous variables (`STATO_0000251`) as specified input. The input data are scalar measurement data (`IAO_0000032`) that are part of (`BFO_0000051`) the continuous variables.

Hence, rather than merely representing the numerical p-value, the approach pursued here describes the performed statistical hypothesis test in a comprehensive and semantic manner, including meaningfully described test input and output. Furthermore, the resulting description is machine readable. The description is an instance of machine readable factual semantic information communicated in scholarly literature.

Given such machine readable descriptions of statistical hypothesis tests e.g., the others included in the paper by Haddad et al. and potentially the many more found in the scientific literature, it is trivial to formulate queries only for statistically significant (specifically, $P < 0.005$ or $P < 0.001$) tests (of a specific kind) involving a particular dependent variable and continuous variables with at least $N$ measurement data. The scientific information communicated in scholarly literature—here the statement that "IRE binding activity was significantly reduced in failing hearts," or more accurately the statistical hypothesis test underlying this statement, with the supporting figures and data in Figure 1B—is thus not just reported in a form suitable for human experts but also available in machine readable form for automated processing.

Technically, the machine readable description of the statistical hypothesis test is a (small) RDF graph, consisting of a set of RDF triples (109 in our example). Various kinds of databases can be used to persist such triples. The most obvious kind is one of the many available triple stores. However, we are

**Listing 2.** Machine readable description of the performed statistical hypothesis test, in RDF Turtle syntax. For the sake of brevity, we omit prefixes but include human readable comments to guide readers through the non-semantic names of OBO Foundry ontology concepts and relations.

```
# a two sample t-test with unequal variance
[] a obo:STATO_0000304 ;
  # that has specified input
  obo:OBI_0000293 [
    # a study design dependent variable
    a obo:OBI_0000751,
      # specifically, iron-responsive element binding
      obo:GO_0030350
  ] ;
  # and has specified output
  obo:OBI_0000299 [
    # a p-value
    a obo:OBI_0000175 ;
    # that has value specification
    obo:OBI_0001938 [
      # a scalar value specification
      a obo:OBI_0001931 ;
      # that has specified numeric value
      obo:OBI_0001937 1.311125e-08
    ]
  ] .
```

currently experimenting with a more general purpose graph database, specifically Neo4j (neo4j.com). The primary motivation for this choice is the possibility, in Neo4j, to attach arbitrary attributes to graph nodes and edges. We plan to make use of this feature to e.g., timestamp data and support versioning.

Aligned with RDF, at the core of our data model is the statement i.e., a structure of three elements (subject, predicate, object) whereby the subject is a resource and the object is either a resource or a literal (predicate is an additional type). Statements, resources, and predicates are identified by means of an internal identifier. With RDF data, URIs are thus mapped to internal identifiers and are, in our data model, the labels of resources or predicates.

A REST API enables interaction with the graph database. Of primary focus here, the API supports the creation and lookup of resources, predicates and statements. Given the RDF triples for the machine readable description of the statistical hypothesis test (Listing 2), we thus implement the storing of triples as statements. Contrary to conventional triple stores, we first need to resolve URIs in triple subject, predicate, and object positions to internal identifiers. Hence, before a statement is stored we perform lookups and create new resources and a predicate in case the corresponding URIs cannot be found (for more detail, see [38]). Given internal identifiers for subject, predicate and resource object we then store the statement. Literal objects are unidentified values.

## 5 Discussion

As suggested by Mons and Velterop for their paper [31], also this paper may appear paradoxical since "it is a paper in classical format that seems to make a plea for the ending of precisely such textual classical publication." Except that this paper is no plea for the ending of classical publication. Rather, we argue that with relatively minor changes to current research infrastructures we may achieve the co-existence of classical publication with machine readable representations of (some of) the information communicated in classical publication.

We suggest that a key element is the prospective (*a priori*) systematic acquisition of machine readable scientific information communicated in scholarly literature i.e., acquisition while researchers perform data analysis and develop the results that build the foundation for the prospective article. This stands in contrast with the (complementary) approach whereby machine readable scientific information is extracted retrospectively (*a posteriori*) from published articles, principally using text mining, possibly combined with human curation.

As shown with our use case, the prospective approach has the potential to capture scientific information at the granularity of individual statements or even numbers reported in tables and figures. We argue that, with current technologies, such granularity cannot be achieved by the retrospective approach, using text mining.

However, the prospective approach relies on changes to the research infrastructure used for data analysis. The challenges are both technical and social. The technical infrastructure needs to be advanced so that the output of computational environments are no longer mere numbers. Rather, numbers need to be information objects with machine readable serialization that captures meaning. Furthermore, the technical infrastructure needs to automatically track relations between entities e.g., to record provenance.

Infrastructure is invisible [35] and this is precisely how the additional functionality delineated here should appear to researchers: invisible. However, some changes in practice are difficult to avoid. Moving data analysis from local computing environments onto interoperable infrastructure e.g., into VREs that interoperate with data and computing resources, is a major change to how data analysis is currently performed, by many if not most researchers and especially those working with little data. Data analysis on local computing environments (e.g., the researcher's workstation) is a key reason for the staggering syntactic and semantic heterogeneity of derivative data generated by researchers in data analysis. In such environments it is hard to harmonize data representation, introduce novel approaches and promote interoperability. Furthermore, the infrastructural discontinuity between local computing environments and engineered research infrastructures makes it difficult or impossible for the latter to monitor workflows and thus track executed activities, retain information about the involved primary and derivative data, as well as to systematically acquire derivative data. Indeed, the download of data from research infrastructures e.g., data repositories is "considered harmful" in most cases [4]. Implications in disciplines with sensitive, personal data such as life sciences need to be considered.

While moving data analysis onto interoperable infrastructure is surely a major social challenge for many research groups and communities, the perspective of performing more of data analysis in well-engineered VREs has great potential as an approach to start addressing the issues discussed here. Naturally, "big science" and "big data" research communities have taken steps into such direction. For example, with CERN Analysis Preservation [14] the High-Energy Physics community is systematically preserving research objects (e.g., data, software) created in analysis. However, the long tail of research with "small data" has arguably been left behind.

The proposed approach can be discussed from the perspective of the FAIR principles for scientific data management and stewardship [44]. The content of Listing 2 is of course data. As they encode scientific information communicated in scholarly literature, the data in Listing 2 are, however, of a kind different from observational data (e.g., sensor network sourced), experimental data (e.g., assay sourced) or computational data (e.g., simulation sourced). In contrast to the form in the article by Haddad et al. (in Figure 1B and in the main text of the article) the data in Listing 2 are clearly more (machine) interoperable. Indeed, the data meet the three requirements for interoperability suggested by the FAIR principles. Specifically, in the proposed form the data are more interoperable because they "use a formal, accessible, shared, and broadly applicable language for knowledge representation"; they "use vocabularies that follow FAIR principles"; and they "include qualified references to other (meta)data." With systematic acquisition in research infrastructures, the proposed approach also supports the findability, accessibility and reusability of scientific information published in scholarly literature, and hence improves on the other elements of the FAIR principles.

The reference to concepts e.g., two sample t-test with unequal variance (`STATO_0000304`) and their formal semantics by means of global and unambiguous identifiers is a key aspect of the FAIR principles. In the proposed approach, infrastructure adopts identified concepts of existing ontologies. The semantics of the resulting data (Listing 2) are thus accessible to machines. This stands in contrast with the natural language text of the original study in which the authors did not make use of ontology concepts.

We implement the proposed approach in Python. With the `rdflib`[5] library, the language has good support for RDF. It is thus straightforward to implement the proposed features in Python. Jupyter supports numerous languages, including R which is another language popular in data science. The effort required to implement the proposed approach in Jupyter but for another language thus depends primarily on whether or not there exists a corresponding RDF library. More flexible approaches may be engineered.

[5] `https://rdflib.readthedocs.io/` (Accessed: October 16, 2018)

Listing 2 only shows iron-responsive element binding (GO_0030350) and the p-value as statistical hypothesis test input and output, respectively. The published Jupyter notebook [38] also includes the data as specified test input. In principle, this description can be extended with further attributes. However, such extension relies on additional vocabulary, likely of a different ontology. For instance, it may be interesting for applications to explicitly capture data summaries e.g., the sample size or the share of NaN values. Such indicators are important in data and statistical test quality assessment. Furthermore, we may capture additional medical context (e.g., ICD-11 codes). To be useful, it is essential for descriptions to adequately capture context. So far, we have given this aspect only limited attention.

## 6 Future work

Though some of the foundations for the infrastructure depicted here have been laid in other disciplinary contexts, specifically the earth and environmental sciences [41, 36, 37, 40, 39, 42], the presented work remains in an embryonic stage. Most of the work required to make the vision [5, 6, 20] reality surely lays ahead. We present here a few avenues for future work.

The application of the approaches originally developed in use cases in earth and environmental sciences to life sciences is important and we are committed to build on the results reported here and develop a compelling use case together with Hannover Medical School as a research infrastructure in life sciences. Such collaboration is essential to determine the requirements for a viable infrastructure.

There exist numerous pathways along which machine readable scientific information can be acquired. In this paper, our focus is on the prospective pathway with data analysis. Also in the category of prospective pathways, we will explore the possibility of acquiring machine readable scientific information at the time of writing the article. Here, it is possible to link existing information objects created e.g., during data analysis with the article. We will explore collaboration with projects such as Dokieli [13] and other document authoring systems.

The retrospective pathways form a further category. They assume one or more written articles, extract scientific information from them, and represent information in machine readable form. In addition to text mining articles, it is interesting to explore the acquisition of machine readable scientific information at the time of article submission. This could be achieved in collaboration with submission systems, such as EasyChair. In addition to metadata about the article, such systems increasingly capture other information e.g., ORCID iDs and funding data. While it is of course untenable to expect a complete "semantification" of the article by the researcher at this point, it is arguably possible to present researchers with a form that captures the key aspects of the research contribution. Text mining could support researchers with suggestions.

As an open project, TIB encourages active stakeholder participation. The project's workshop series is a key instrument to this effect. We invite domain scientists to contribute requirements, use cases and domain expertise; representatives of related projects such as FREYA, OpenAIRE, Research Graph to explore synergies among infrastructures; representatives of the publishing sector (articles, data and other artefacts) for their related work and possible future integrations.

## 7 Conclusion

For a use case in life sciences, we have demonstrated how research infrastructures can systematically acquire machine readable scientific information communicated in scholarly literature. We argue that this possibility is enabled by the technological integration of VREs (in particular components for data analysis) and semantic technologies. While technical challenges do exist, we argue that the greater challenges are social, specifically the required changes in research practices. Indeed, data analysis currently performed on local computing environments needs to move into VREs. Such environments can be engineered to include novel functionality that enables the systematic acquisition of scientific information so that information is also represented in machine readable form using technologies that not only represent data but also their meanings.

**Acknowledgements.** We thank the TIB Leibniz Information Centre for Science and Technology for supporting this project and our colleagues and the participants of the project's workshop series for their contributions.

## References

1. A. Aamodt and M. Nygård. Different roles and mutual dependencies of data, information, and knowledge – An AI perspective on their integration. *Data & Knowledge Engineering*, 16(3):191–222, 1995.
2. R. Allan. *Virtual Research Environments: From Portals to Science Gateways*. Chandos Publishing, Oxford, 2009.
3. A. Aryani and J. Wang. Research Graph: Building a Distributed Graph of Scholarly Works using Research Data Switchboard. In *Open Repositories CONFERENCE*, 2017.
4. M. Atkinson, R. Filgueira, A. Spinuso, and L. Trani. Download considered harmful. 2018. Manuscript in preparation.
5. S. Auer. Towards an Open Research Knowledge Graph, Jan. 2018.
6. S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, and M. E. Vidal. Towards a Knowledge Graph for Science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, WIMS '18, pages 1:1–1:6, New York, NY, USA, 2018. ACM.
7. J. Barwise and J. Perry. Situations and Attitudes. *The Journal of Philosophy*, 78(11):668–691, November 1981.
8. S. Bechhofer, D. D. Roure, M. Gamble, C. Goble, and I. Buchan. Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*, jul 2010.

9. L. Bornmann and R. Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015.
10. A. Burton, A. Aryani, H. Koers, P. Manghi, S. L. Bruzzo, M. Stocker, M. Diepenbroek, U. Schindler, and M. Fenner. The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Magazine*, 23(1/2), jan 2017.
11. L. Candela, D. Castelli, and P. Pagano. D4Science: an e-Infrastructure for Supporting Virtual Research Environments. In M. Agosti, F. Esposito, and C. Thanos, editors, *Proceedings of the 5th Italian Research Conference on Digital Libraries (IRCDL 2009)*, Padova, Italy, January 2009.
12. L. Candela, D. Castelli, and P. Pagano. Virtual Research Environments: An Overview and a Research Agenda. *Data Science Journal*, 12(0):GRDI75–GRDI81, 2013.
13. S. Capadisli, A. Guy, R. Verborgh, C. Lange, S. Auer, and T. Berners-Lee. Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli. In *International Conference on Web Engineering*, pages 469–481, 2017.
14. X. Chen, S. Dallmeier-Tiessen, A. Dani, R. Dasler, J. D. Fernández, P. Fokianos, P. Herterich, and T. Šimko. CERN Analysis Preservation: A Novel Digital Library Service to Enable Reusable and Reproducible Research. In N. Fuhr, L. Kovács, T. Risse, and W. Nejdl, editors, *Research and Advanced Technology for Digital Libraries*, pages 347–356, Cham, 2016. Springer International Publishing.
15. P. Ciccarese, M. Ocana, L. J. Garcia Castro, S. Das, and T. Clark. An open annotation ontology for science on web 3.0. *Journal of Biomedical Semantics*, 2(2):S4, May 2011.
16. H. V. de Sompel, S. Payette, J. Erickson, C. Lagoze, and S. Warner. Rethinking Scholarly Communication. *D-Lib Magazine*, 10(9), sep 2004.
17. A. de Waard, L. Breure, J. G. Kircz, and H. van Oostendorp. Modeling Rhetoric in Scientific Publications. In *Proceedings of the International Conference on Multidisciplinary Information Sciences and Technologies (InSciT2006)*, 2006.
18. A. de Waard, S. B. Shum, A. Carusi, J. Park, M. Samwald, and Á. Sándor. Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims. In T. Clark, J. S. Luciano, M. S. Marshall, E. Prud'hommeaux, and S. Stephens, editors, *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, volume 523, Washington DC, USA, October 2009. CEUR.
19. K. Devlin. *Logic and Information.* Cambridge University Press, 1991.
20. S. Fathalla, S. Vahdati, S. Auer, and C. Lange. Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles. In J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, and I. Karydis, editors, *Research and Advanced Technology for Digital Libraries*, pages 315–327, Cham, 2017. Springer International Publishing.
21. L. Floridi. *The Philosophy of Information.* Oxford University Press, 2011.
22. L. J. García-Castro, O. X. Giraldo, and A. García-Castro. Using annotations to model discourse: an extension to the Annotation Ontology. In A. García-Castro, C. Lange, F. van Harmelen, and B. Good, editors, *Proceedings of the 2nd Workshop on Semantic Publishing*, volume 903, pages 13–22, Hersonissos, Crete, Greece, May 2012. CEUR.
23. P. Groth, A. Gibson, and J. Velterop. The anatomy of a nanopublication. *Information Services & Use*, 30(1-2):51–56, sep 2010.

24. S. Haddad, Y. Wang, B. Galy, M. Korf-Klingebiel, V. Hirsch, A. M. Baru, F. Rostami, M. R. Reboll, J. Heineke, U. Flögel, S. Groos, A. Renner, K. Toischer, F. Zimmermann, S. Engeli, J. Jordan, J. Bauersachs, M. W. Hentze, K. C. Wollert, and T. Kempf. Iron-regulatory proteins secure iron availability in cardiomyocytes to prevent heart failure. *European Heart Journal*, 38(5):362–372, 2017.
25. K. L. Hanson, T. DiLauro, and M. Donoghue. The RMap Project: Capturing and Preserving Associations Amongst Multi-Part Distributed Publications. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '15, pages 281–282, New York, NY, USA, 2015. ACM.
26. M. W. Hentze, M. U. Muckenthaler, B. Galy, and C. Camaschella. Two to Tango: Regulation of Mammalian Iron Metabolism. *Cell*, 142(1):24–38, jul 2010.
27. A. E. Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.
28. E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.
29. T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, and J. development team. Jupyter Notebooks—a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90. IOS Press, 2016.
30. F. Manola, E. Miller, and B. McBride. RDF Primer. Recommendation, W3C, February 2004.
31. B. Mons and J. Velterop. Nano-publication in the e-science era. In *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, Washington, DC, USA, 2009.
32. J. Priem. Beyond the paper. *Nature*, 495(7442):437–440, mar 2013.
33. C. A. Schneider, W. S. Rasband, and K. W. Eliceiri. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7):671–675, jul 2012.
34. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, T. O. Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, nov 2007.
35. S. L. Star. The ethnography of infrastructure. *American Behavioral Scientist*, 43(3):377–391, 1999.
36. M. Stocker. Advancing the Software Systems of Environmental Knowledge Infrastructures. In A. Chabbi and H. W. Loescher, editors, *Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities*, pages 399–423. Taylor & Francis Group, CRC Press, 2017.
37. M. Stocker. From Data to Machine Readable Information Aggregated in Research Objects. *D-Lib Magazine*, 23(1/2), jan 2017.
38. M. Stocker. Jupyter notebook for DILS 2018 paper on research infrastructures that curate scientific information. figshare, July 2018.
39. M. Stocker, E. Baranizadeh, H. Portin, M. Komppula, M. Rönkkö, A. Hamed, A. Virtanen, K. Lehtinen, A. Laaksonen, and M. Kolehmainen. Representing situational knowledge acquired from sensor data for atmospheric phenomena. *Environmental Modelling & Software*, 58:27–47, 2014.
40. M. Stocker, J. Nikander, H. Huitu, M. Jalli, M. Koistinen, M. Rönkkö, and M. Kolehmainen. Representing Situational Knowledge for Disease Outbreaks in Agriculture. *Journal of Agricultural Informatics*, aug 2016.

41. M. Stocker, P. Paasonen, M. Fiebig, M. A. Zaidan, and A. Hardisty. Curating Scientific Information in Knowledge Infrastructures. *Data Science Journal*, 17, 2018.
42. M. Stocker, M. Rönkkö, and M. Kolehmainen. Situational knowledge representation for traffic observed by a pavement vibration sensor network. *IEEE Transactions on Intelligent Transportation Systems*, 15(4):1441–1450, August 2014.
43. K. E. White, C. Robbins, B. Khan, and C. Freyman. Science and Engineering Publication Output Trends: 2014 Shows Rise of Developing Country Output while Developed Countries Dominate Highly Cited Publications. Technical Report NSF 18-300, National Science Foundation, Oct. 2017.
44. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, mar 2016.