Semantic-Aware Fingerprints in RDF Metadata and Resilience Research

Hans-Gert Gräbe[1][0000−0002−3934−413X]

InfAI, Leipzig University, Leipzig, Germany
graebe@infai.org

Abstract. With the concept of resilience, an attempt is made in the context of socio-cultural, socio-economic and socio-ecological systems to identify conditions for stabilising development paths and to implement these in practical management. In doing so, short-term cost-benefit calculations and long-term effects are often in contradiction with each other. Resolving such contradictions is difficult. Data-driven resilience research offers a way to articulate more precisely these contradictions, whose dynamics often span different spatio-temporal levels. Accordingly, data structure concepts are needed to embed domain-specific semantics into cross-domain systemic structures. Such means are necessary to coordinate the resilient management of individual resources and that of entire resource pools. The concept of semantic-aware fingerprints presented here can contribute to this.

Keywords: Semantic-aware fingerprints · Resilience · RDF metadata · Converting metadata to RDF.

1 Introduction

The concept of resilience was first developed in psychology in the 1950s to address questions of the robustness and adaptability of individuals to changing socio-cultural conditions [9]. Since at least the 1970s, the concept has also been applied to questions of ecosystem adaptability [10]. In contrast to psychological approaches, which primarily address the conditions of possible development under *established* personal individual structures, the study of the adaptability of ecosystems are focused on, if not centered at, formation and design of appropriate development conditions. This is not surprising, since existing ecosystems have been socio-culturally shaped by human activity for thousands of years and there are hardly any "natural" ecosystems left on our planet. Accordingly, not only descriptive and explanatory approaches play a role in this research, but also modelling, planning and implementation aimed at redesigning ecosystems towards greater resilience. A distinction is often made between adaptive and transitional management approaches [2, 6].

Such research is based on systemic concepts of the delimitation and transformation of ecosystems [11], whereby the transformation needs are often closely

linked to problem situations that manifest themselves in *contradictory development perspectives*, as examined in more detail in [6].

Development perspectives charged with design claims are always linked to *socio-cultural goals*, whereby the most fundamental contradictions often manifest themselves as those *between short term and long term goals*. Such contraditions cannot be resolved based on simple systemic modelling, since systemic modelling is always *reductionistic* [6] and the necessary delimitation is determined by systemic eigentimes and eigenspaces [11]. To describe approaches of *coupling* developmental dynamics of such systems which evolve in different spatio-temporal dimensions as a *system of systems*, C.S. Holling proposed in [11] the concept of a *panarchy*, see also [6].

Data-driven resilience research is thus faced with the task of contributing to the resolution of those fundamental contradictions between short term and long term goals. These contradictions can only be insufficiently addressed by existing approaches [3, 11, 6]. In this context, *metadata management* is of particular importance as an element of the coupling between the levels of the dynamics of individual resources and the dynamics of the management of a resource pool [7]. With the concept of *semantic-aware fingerprints* we discuss the question of semantic transport in such couplings.

2 Converting Metadata to RDF

In addition to its use in the agreement and standardisation of conceptual systems as a socio-cultural process, the *Resource Description Framework* (RDF) plays an important role also as a universal metadata format, especially for the brief presentation of data sets as well as for the search and retrieval of concrete data as the *resources* described in the RDF metadata record.

When extracting corresponding metadata from existing datasets or transforming them to RDF, there is always the question of whether more complex substructures, such as e.g. Geodata formats, should be transformed or whether it is better to keep them in the given domain-specific conceptualisations and serialisation formats, since such domain specific representations are often both optimised in terms of storage space and there exist already sufficient powerful tools for their visualisation and processing based on that domain-specific serialisation format.

Transformation to RDF and use of such data faces two main problems:

1. Required effort and losses during transformation are sometimes high. Especially the restrictions resulting from the concept of RDF data as *sets* of three-word sentences often do not support a representation of sequential and operational relationships in the data.
2. There are no tools for the transformed data that are comparable in their performance with those from the domain.

On the other hand, the use of original data formats from the domain makes cross-domain search processes considerably more difficult. Indexing using classic

hash functions accesses only syntactic structures and thus cannot express semantic equivalence of syntactically different data. The representation of semantic aspects in metadata therefore requires the collaboration of domain-specific expertise and the semantic representation of selected domain-specific concepts at the meta-level, for example in the form of RDF predicates.

Since metadata is closely related to the *practice* of using the data itself also beyond the given domain of expertise, and this in turn to the *knowledge* of selected domain-specific concepts as well as the *use* of selected standardised domain-specific tools, such selected domain-specific concepts are also required for the general user to be aware of.

At this point we note once again that Semantic Web technologies are less concerned with the production of a Linked Open Data Cloud as an *artefact* than with a process of *cooperative action* of developing and consolidating common conceptual systems as a *prerequisite* for the decentralised collection and use of semantically significant data. The focus is thus not so much on the data collection as the *result*, but rather on the complex cooperative socio-technical *process* of collecting the data as a *mental activity* (mysledeyatel'nost') in the sense of Shchedrovitsky [17].

3 An Example

We first encountered this question "to convert or not to convert" with the PoSSo project [16], where, after its end in 1995, we [18] were concerned with compiling the collected benchmark problems for solving polynomial systems in a reliable form. During the PoSSo project this data had been collected and stored on various computers of the project partners or were available even in printed form only. In designing a markup format suitable for that purpose, we were faced also with the question of a transformation of the polynomial notation commonly used in mathematics in our own markup. Already at that time some early adopters of MathML [14] or OpenMath [15] strongly argued in that direction. Although these formats, which were standardised later on, allow an exact specification of commonly used mathematical function and operator symbols, they also lead to a significant blow-up of the data size. Even more, when modern CAS such as *Mathematica* or *Maple* read in MathML inputs today, they are first transformed into the "usual" mathematical format (more precisely: into the CAS-internal representation of this format) to continue working with it.

The decision in the SymbolicData project was to use XML-like structures[1] to delimit various metadata up to the representation of the lists of polynomial, but to leave the polynomials themselves in their usual mathematical notation of a *distributive normal form*, see [5] for details.

This already brings us to semantic awareness, because the decision to represent polynomials in this or, say, a MathML notation is preceded by the decision

[1] RDF did not yet exist at that time, in 1998, XML was not yet sufficiently standardised.

to use that distributive normal form, a technical term deeply rooted in the domain, well known to domain experts, but less to a general user.

For those general users, understanding the general term "polynomial" may be sufficient for using the dataset. However, if searching and finding in this data is to be organised in a way that involves the polynomial systems themselves, these domain terms are essential in their *semantics*. However, general query systems like SPARQL are not designed for this and are also difficult to extend with corresponding domain-specific concepts.

Even though the distributive normal form provides a canonical form and thus guarantees syntactic uniqueness of the representation of polynomial systems under certain restrictions, these conditions were not given in our use case: We were concerned with identifying examples that in different form represent the same polynomial system. Such examples could differ using other variable names or other variable orders. To decide matching completely would have required complex and time-consuming calculations, even though these in principle could be automated with the tools existing in the *Computer Algebra* domain.

Instead, we decided to look for *semantic-aware fingerprints* of such polynomial systems, i.e. invariants that are easily to be computed as well as easily stored and searched, but achieve a high, though not necessarily complete, discriminatory power on the given collection of polynomial systems. As such fingerprints were used: the number of variables, the set of numbers of terms per polynomial and the set of degrees of polynomials (both realised as ordered lists of integers). Problems arising from the partial lack of full discriminatory power of the fingerprints were resolved by closer inspection of the examples themselves that could not be distinguished. In all cases it was sufficient to inspect the respective scientific context, since it was known and explained why the examples only slightly differ (for example, because one example had emerged from the other through a misprint).

For the expansion of the collection with new examples, a closer inspection of the candidates is necessary, but again only against examples in the collection with the same fingerprint. This drastically reduces the required domain-specific workload.

4 Domain Specific Data, Indexing and Metadata

The example explained in the previous section is typical for metadata management challenges to organise domain-specific data for a wider audience, the design of management, search and filter functionality. For this purpose data is usually *indexed* based on metadata that collect important relevant information of the individual data records in a compact manner. As in the previous section we denote such metadata for an individual data record as its *fingerprint*.

Similar to a hash function a fingerprint function computes a compact metadata record (a *resource description* in the RDF terminology) to each individual data record (*resource* in the RDF terminology). As with a hash function one can use the fingerprints to (almost) distinguish different data records within

the given collection and to match new records with given ones. But there is an essential difference between (classical) hash functions and well designed fingerprints: fingerprint functions exploit not only the textual representation of the data record as meaningless syntactical character string but convey semantically important information based on domain-specific concepts or even compute such information from the resource using domain-specific tools. Fingerprints are in this sense *semantic-aware* and can even be designed in such a way that they map ambiguities in the textual representation of records (e.g., polynomial systems given in different variable orders and even in different variable sets, as explained in the previous section) to *semantic invariants*.

The design of appropriate fingerprint signatures is an important *intracommunity* activity to structure data collections. Such fingerprint signatures are also very useful for the *intercommunity* usage if the data is provided by the domain specific community to a wider audience as a *service*, since they allow to navigate within the (foreign) data collection without presupposing the full knowledge of the "general nonsense" of the given domain, i.e., the informal background knowledge required to be known to a specialist in that domain. Hence well designed fingerprint signatures are to be considered also as a first class service of a domain-specific community to a wider audience to inspect their data collections without using the domain-specific tools to access the resources themselves.

5 Working with Semantic-Aware Fingerprints

Usually data collections of a certain community are stored in a specially designed community-internal format, often as plain text, in a special XML notation or as SQL database. Such formats usually employ special formal semantics agreed within the community as an effective way to store domain specific input and output data and used by commonly developed tools with appropriate parsing functionality.

Usually such formats are extended to store metadata, i.e. fingerprints, together with the data in a single resource as, e.g., in the IEEE Learning Object Metadata (LOM) Standard [12]. This has one benefit and two drawbacks:

- *Benefit:* A fingerprint can be computed immediately by the commonly used tools or with their slight extension, and can be stored with the resource itself.
- *First Drawback:* Metadata unfold its full expressiveness only if it can be searched and navigated. Storing metadata together with the resource itself implies high extraction costs for navigation and access to the data collection as a whole.
- *Second Drawback:* The very different formats prevent an easy combination of metadata from different communities and even from different sources.

The first drawback can be addressed if the metadata are extracted into a database accompanying the data collection and provide *intracommunity tools* for search and navigation within that metadata. Such an approach based on a web interface was realised, e.g., within the ELMAT project [1]. The metadata is stored

in a database and is available only within OPAL – the Saxonian E-Learning Platform – as intracommunity tool.

Such a solution has two further drawbacks:

- The search and navigational functionality is not or only in a restricted way adapted for machine-readable *interaction* and thus cannot be integrated into more comprehensive search and navigational processes.
- The search and navigational functionality can't be adapted by the user for its own needs.

A well known general solution that avoids these drawbacks proposes to extract the metadata information from the resource data, to transform it into RDF and thus to make it available for interlinking within the *Linked Open Data World* as a worldwide distributed database that can be globally queried and navigated using the SPARQL query language in a similar unified way as SQL allows to navigate in local relational databases.

Semantic-aware fingerprints are an important tool to anchor *domain-specific semantics* in such an overarching search process. The question to determine more precisely which domain-specific concepts and to what degree of detail are relevant for further application can only be clarified in a discoursive negotioation process in which data provider and data user act on equal level. Only in such an *organisational* framework of resource *management* of data stocks the stable availability of up-to-date data sources can be organised, which in turn form the basis for not only qualitative but also *quantitative change management* and thus provide the linguistic means to base the topic of resilience on a data-driven and thus scientific foundation.

In this context, action and negotiation are closely related: the practical *creation* and management of domain-specific data stocks in the context of *domain-specific inner logics* and the *outer logics* of the *use* of these data stocks in other contexts with other domain-specific inner logics initially manifest themselves in the concurrent, parallel action of several subsystems and must be condensed into a new overarching systemic context through negotiation, as explained in more detail in [7, 8].

6 Conclusion

In this paper, we have mainly touched *structural* issues of building metadata vocabularies to anchor domain-specific semantics appropriately at a cross-domain level. However, this is solely the *substrate* in the sense as Goodwin [4, p. 38] uses this notion on which cooperative action unfolds at both systemic levels that are connected via the coupling of resources and metadata.

We have shown that semantic awareness at the meta-level through fingerprints is well suited to localise or even identify problematic resources based on suitable parameters. The concept of semantic-aware fingerprints can thus be well integrated into *Systematic Innovation Methodologies* such as TRIZ [13], which are not based on pure brainstorming and trial-error concepts as a number

of adaptive methodologies, but pursue clear transitional concepts and rely on concise modelling, Ideal Final Results, identification of core contradictions and focused problem solving based on this.

Nevertheless *both* approaches are attempts to understand the larger whole from its parts and to operate that whole from such an understanding. Goodwin [4] draws attention to the fact that such a "system concept of first kind" (Shchedrovitsky [17, p. 91]) is at best the *substrate* for the living dynamics of cooperative action and that the stratification of the materiality and operational forms of living systems must be grasped differently, with a "system concept of second kind" (Shchedrovitsky). Since "a living system has no parts" (ibid.).

References

1. BPS Sachsen GmbH: ELMAT – Elektronische Übungs- und Bewertungstools für Mathematikveranstaltungen (in German). Gefördert vom Sächsischen Staatsministerium für Wissenschaft und Kunst und dem Arbeitskreis E-Learning bei der Landesrektorenkonferenz Sachsen.
2. Foxon, T.J., Reed, M.S., Stringer, L.C.: Governing long-term social–ecological change: what can the adaptive management and transition management approaches learn from each other? In: Environmental Policy and Governance, 19 (1), 2009, pp. 3–20. doi:10.1002/eet.496
3. Geels, F.W. Schot, J.: Typology of Sociotechnical Transition Pathways. In: Research Policy 36 (2007), pp. 399–417. doi:10.1016/j.respol.2007.01.003
4. Goodwin, C.: (2018). Co-operative Action. Cambridge University Press, 2018. ISBN 978-1-108-71477-8.
5. Gräbe, H.-G.: Semantic-aware Fingerprints of Symbolic Research Data. In: Greuel, G.-M., Koch, T., Paule, P., Sommese, A. (eds.) Mathematical Software – ICMS 2016. LNCS, vol. 9725, pp. 411–418. Springer, Heidelberg (2016). doi:10.1007/978-3-319-42432-3.
6. Gräbe, H.-G., Kleemann, K.P.: Seminar Systemtheorie (in German). Rohrbacher Manuskripte, Heft 22. Berlin, 2020. ISBN 9783752620023.
7. Gräbe, H.-G.: Systems and systemic development in TRIZ. Submitted to the TRIZ Future Conference 2022.
8. Gräbe, H.-G.: Components as Resources and Cooperative Action. Submitted to *Deutscher TRIZ-Anwendertag* 2022.
9. Haas, M.: Stark wie ein Phönix (in German). OW Barth, 2015.
10. Holling, C.S.: Resilience and stability of ecological systems. In: Annual Review of Ecology and Systematics vol. 4, 1973, pp. 1–23.
11. Holling, C.S.: Understanding the Complexity of Economic, Ecological, and Social Systems. In: Ecosystems (2001) 4, pp. 390–405.
12. IEEE Standard for Learning Object Metadata. IEEE Std 1484.12.1™-2020.
13. Mann, D.: Hands-On Systematic Innovation for Business and Management. IFR Press, 2007.
14. Mathematical Markup Language (MathML). Version 3.0, 2nd Edition. W3C Recommendation 10 April 2014.
15. The Open Math Project. https://openmath.org/
16. The PoSSo Project. Polynomial Systems Solving – ESPRIT III BRA 6846. (1992–1995). https://cordis.europa.eu/project/id/6846. [202-05-05]

17. Shchedrovitsky, G. P.: Selected Works. A Guide to the Methodology of Organisation, Leadership and Management. In: Khristenko, V.B., Reus, A.G., Zinchenko, A.P. et al.: Methodological School of Management. Bloomsbury Publishing (2014).
18. The SymbolicData Project. (1998–2018). https://symbolicdata.github.io