Semantic Web for the Humanities

Albert Meroño-Peñuela
Advisors: Stefan Schlobach, Frank van Harmelen

Department of Computer Science, VU University Amsterdam, NL
albert.merono@vu.nl

Abstract. Researchers have been interested recently in publishing and linking Humanities datasets following Linked Data principles. This has given rise to some issues that complicate the semantic modelling, comparison, combination and longitudinal analysis of these datasets. In this research proposal we discuss three of these issues: representation round-tripping, concept drift, and contextual knowledge. We advocate an integrated approach to solve them, and present some preliminary results.

Keywords: Semantic Web, Formats, Concept Drift, Contexts

1 Motivation and research questions

Humanities researchers have been interested recently in publishing and linking their datasets following Linked Data principles, in order to enhance their decentralization, openness, changeability and integration. Traditionally, the unique demands of the Humanities, their limited technical and modelling interests, and the highly contextualized nature of their source materials have kept this field distant from the Semantic Web.

We make efforts to bridge the gap. As case studies, we convert Dutch historical censuses (1795-1971) and the catalogue of publications in the Netherlands during the Golden Age (STCN, 16th century onwards) to RDF [10], we model them using standard vocabularies, and we publish them on the Web.

These datasets are messy and heterogeneous. Different dataset versions contain inconsistent structuring rules, concepts with a changing meaning over time, and multiple representation formats. Comparison, combination and longitudinal queries (e.g. *evolution on the number of shoemakers in Amsterdam from 1795 to 1971*) are notoriously difficult. Researchers are forced to manually rewrite data and queries, incurring in high labour costs and non repeatable practices.

Figure 1 shows data heterogeneity interacting with other indicators. Since our goal is to increase data integration, data heterogeneity has to lower, as shown by arrows and signs. Lowering data heterogeneity is no trivial task, and we identify *format round-tripping*, *concept drift* and *contextual knowledge* as influencing indicators that can indirectly improve data integration.

Format round-tripping. Lots of data formats are used to encode semistructured datasets. Tools for legacy conversion between these formats are required: Humanities researchers use non RDF compatible tools, and providing data in

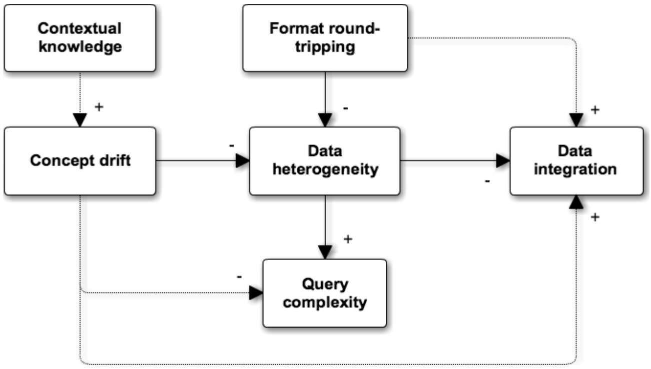

Fig. 1: Indicators influence each other as indicated by arrows. The increase of any given indicator increases/decreases, respectively, the one influenced by it as indicated by the +/- signs. E.g., increasing format round-tripping decreases data heterogeneity.

various formats on demand is a requirement. Under this topic we will investigate, first, how to perform any plain, tabular, tree-based, graph-based or relational-based format conversion from a holistic point of view and, second, whether the original data can be retrieved after arbitrary conversions.

Concept drift. Different versions of the same dataset show that concepts change their meaning over time, especially if the time gap is wide. Although not meaning exactly the same, two time gapped instances of the same concept may preserve some degree of sameness. For example, the concept *shoemaker* in the 17th century (someone who makes shoes with leather) has drifted until nowadays (someone who owns a company). Mapping drifted concepts correctly is necessary to solve longitudinal queries in Humanities data.

Contextual knowledge. Humanities ontologies require dynamic concept formalizations instead of static ones, especially for contested, open-textured or ambiguous concepts. The definition of such concepts needs to be dynamically built depending on their contexts. Examples of contextual knowledge are the time when and the space where the concept occurs, subjective opinions on the concept, or domain expert statements about the concept. Multiple contradictory definitions may need to coexist in one ontology.

Concept drift and knowledge from contexts are closely related. Since the context of a concept often changes over time, a definition of concept drift based on the varying properties in contexts can be established. Despite less connected, formats often define metadata describing dataset contextual information, which needs to be appropriately modelled. We realize these phenomena are not exclusive to the Humanities, and this proposal looks further on solving longitudinal analyses in dynamic domains of any kind.

We define a general goal of *providing algorithms, formalisms and tools to disambiguate, clean, prepare, normalize, transform, link and query Humanities datasets, conforming a framework for effective Humanities data publishing in the Semantic Web.* Under this umbrella, our research questions are:

1. Can RDF data models faithfully represent the Humanities sources? Is an RDF-based format round-tripping framework possible?
2. How can we model concept drift? Can drifted concepts be aligned?
3. Can we infer dynamic concept definitions from explicitly formalized contexts? Can these contexts help solving concept drift?

2 State of the art

Work has been developed on translating RDF, spreadsheet formats and relational databases. Conversion from relational databases to RDF is covered by [7,13], and the W3C has developed a standard (R2RML) for this purpose. Some tools like D2RQ allow accessing relational databases as virtual RDF graphs. Translating RDF backwards to the relational model is developed in [12] under some assumptions. Conversion between spreadsheet formats and RDF is also possible [6,10]. Google Refine is a power tool for working with messy data and generic format translations, with plugins supporting RDF.

Concept drift in the Semantic Web has been studied in [14], where the authors establish a theory for concept drift defining the meaning of a concept in terms of its intension, extension and labeling. Other Semantic Web approaches have used conceptual clustering [2] or concept signatures [5] to detect concept drift. In Description Logics, ontology diff [3] can be used to determine meaning differences. The question has been discussed in Philosphy around the confrontation of history of unit-ideas versus a pure linguistic intellectual history [9].

Some work has been done recently with respect to contexts in the Semantic Web, although they emphasize the specific goals of improving data integration [4] or speeding up reasoning [11]. Rule interchange languages for the Semantic Web like RIF are also related to dynamic concept construction [8].

3 Proposed approach

Format round-tripping. Existing approaches on format conversion pair any data format with RDF and perform a forward or backward transformation between the two. Our proposal is to take an holistic approach, studying the expressivity of these languages and checking whether arbitrary translation workflows are possible. We are interested in round-tripping translation paths to check if original representations can be regenerated without data loss. We aim at canonical RDF graph forms [1] and centric RDF data representations.

Concept drift. We will study what precise relationship holds between two different versions of a changing concept, identifying the presence of a drift and its nature. Using Description Logics work on ontology diff [3], we will define a minimum meaning concept core, which keeps stable over time despite other non essential transformations. A data model to represent drifted concepts will be needed. A systematic comparison between unstable concept properties will tell whether a drift occurred, and its type. We consider discussions on history of unit-ideas [9] and theories of concept drift for the Semantic Web [14] as inspiration.

Contextual knowledge. We will study how concepts can be dynamically defined depending on their graph contexts. To solve contextual knowledge questions we aim at a two step process. First, we target an explicit semantic representation of the context of a concept, and we will use various data models and vocabularies to define contexts. Second, we consider inference for deriving logical consequences from previously selected contextual graphs. This process can be further integrated with our concept drift framework.

4 Research methodology

We establish an iterative workflow that runs the proposed topics in parallel, first developing theories and then proposals. Proposals will be evaluated with at least the two Humanities case-studies referred in Section 1. All models and automated methods will be validated by domain experts. At the end of each iteration, resulting design methods will be scaled up and refined.

5 Results and future work

Regarding format round-tripping, we implemented some preliminary tools[1,2]. `TabLinker` is a MS Excel to RDF converter supporting translation of annotations and interactive user defined mappings. We also developed scripts generating RDF from various semistructured data formats. We plan to evaluate round-tripping by comparing an original file with its circularly translated homologue. With respect to concept drift, a first set of mappings between possibly label-drifted concepts have been defined using label similarity functions. We run simple longitudinal queries with `MP2Demo`, relying on an hybrid top-down/bottom-up approach that combines upper ontologies (e.g. Historical International Standard Classification of Occupations, HISCO) with automatically extracted local ontologies.

Three more yearly iterations will be carried out. Format round-tripping will be generalized from current scripts, defining transformation entities that will abstract specific format dependencies to modelling artifacts. We will create a data model for concept drift and an RDF/OWL simulation framework to test it with ontology diff and intension, extension and labeling functions. We will extend this framework to integrate reasoning with contexts.

6 Conclusion

In this research proposal we motivate the problems of format round-tripping, concept drift, and contextual knowledge in the context of a Humanities enabled Semantic Web. We propose an approach with novel perspectives extending the state of the art, and we describe an iterative research method to sort these issues

[1] `http://github.com/Data2Semantics/`

[2] `http://github.com/CEDAR-project/`

out. Finally, we show work that has been done during the first year iteration, and we establish a plan for the remainder.

Acknowledgements The work on which this paper is based has been partly supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the CEDAR project. For further information, see http://ehumanities.nl.

References

1. Carroll, J.J.: Signing RDF Graphs. Tech. Rep. HPL-2003-142, HP Lab (2003)
2. Fanizzi, N., d'Amato, C., Esposito, F.: Conceptual Clustering and Its Application to Concept Drift and Novelty Detection. In: The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008. Proceedings. pp. 318–332. ESWC'08, Springer-Verlag, Berlin, Heidelberg (2008)
3. Gonçalves, R.S., Parsia, B., Sattler, U.: Analysing multiple versions of an ontology: A study of the NCI Thesaurus. In: Proceedings of the 24th International Workshop on Description Logics (DL 2011) (2011), http://ceur-ws.org/Vol-745/
4. Guha, R., Mccool, R., Fikes, R.: Contexts for the Semantic Web. In: The Semantic Web - ISWC 2004: Third International Semantic Web Conference. Proceedings. pp. 32–46. Lecture Notes in Computer Science, 3298, Springer (2004)
5. Gulla, J.A., et al.: Semantic Drift in Ontologies. In: Filipe, J., Cordeiro, J. (eds.) WEBIST 2010, Proceedings of the 6th International Conference on Web Information Systems and Technologies. pp. 13–20. INSTICC Press (2010)
6. Han, L., Parr, C., Sachs, J., Joshi, A.: RDF123: a mechanism to transform spreadsheets to RDF. Tech. rep., University of Maryland, Baltimore County (2007)
7. Korotkiy, M., Top, J.L.: From Relational Data to RDFS Models. In: Web Engineering - 4th International Conference (ICWE) Munich, Germany, July 26-30, Proceedings. pp. 430–434. Lecture Notes in Computer Science, 3140, Springer (2004)
8. Krisnadhi, A., Maier, F., Hitzler, P.: OWL and rules. In: Proceedings of the 7th International Conference on Reasoning Web. pp. 382–415. RW'11, Springer (2011)
9. Kuukkanen, J.M.: Making Sense of Conceptual Change. History and Theory 47, 351–372 (2008)
10. Meroño-Peñuela, A., et al.: Linked humanities data: The next frontier? a case-study in historical census data. Proceedings of the 2nd International Workshop on Linked Science (LISC2012). International Semantic Web Conference (ISWC) (2012), http://ceur-ws.org/Vol-951/
11. Peñaloza, R., Baader, F., Knechtel, M.: Context-Dependent Views to Axioms and Consequences of Semantic Web Ontologies. Web Semantics: Science, Services and Agents on the World Wide Web 12(0) (2012)
12. Ramanujam, S., et al.: R2D: Extracting Relational Structure from RDF Stores. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01. pp. 361–366. WI-IAT '09, IEEE Computer Society, Washington, DC, USA (2009)
13. Sequeda, J.F., Arenas, M., Miranker, D.P.: On directly mapping relational databases to RDF and OWL. In: Proceedings of the 21st World Wide Web Conference, WWW 2012. pp. 649–658. WWW '12, ACM, New York, NY, USA (2012)
14. Wang, S., Schlobach, S., Klein, M.C.A.: What Is Concept Drift and How to Measure It? In: Knowledge Engineering and Management by the Masses - 17th International Conference, EKAW 2010. Proceedings. pp. 241–256. Lecutre Notes in Computer Science, 6317, Springer (2010)