Appendix: WHiSe 2020 Diary

Under the circumstances of the COVID-19 pandemic, one of the greatest challenges of WHiSe 2020 was to provide a lively and interactive workshop on an intense, full-day online schedule and, unlike most face-to-face events, with awareness of timezone differences. The core infrastructure and support was provided to us by the ESWC conference organization, mainly in the form of a Web-based system for video streams and chats. This effectively permitted the bold choice of streaming all presentations live from the presenters' homes or workplaces, rather than having them pre-recorded. In addition, and in order to facilitate the flow and interactions between the audience and the presenters while still finishing on time, we provided a link to a Web 'live document' that any participant could edit at any time. This document had been pre-filled with slots for the various workshop sessions and presentations, where the audience were instructed to write down their questions, suggestions and comments as the workshop progressed. The resulting interactions could be summarised as a small, parallel 'digital' track of the workshop with its own set of interesting observations. At the end of each presentation, the session chairs moderated those to be relayed live to the presenters, who were, however, also asked to follow up on the others offline. This led to a level of detail that would have been otherwise unattainable in the limited online setting and timing of the workshop.

As a documented account of the WHiSe response to the pandemic, in what follows we offer a curated, corrected and cleaned version of the minutes of WHiSe 2020. We hope this will serve as a reference of these interesting discussions, as well as an inspiration for both authors and readers for future work.

Workshop Log

Introduction / 10:00 – 10:20

Chair: Alessandro Adamou (@anticitizen7x)

Session 1: Linked Data and Libraries / 10:20 – 12:00

Chair: Albert Meroño-Peñuela (@albertmeronyo)

Presentation: Minna Tamper, Petri Leskinen, Jouni Tuominen and Eero Hyvönen. "Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology".

Albert HISCO is a similar effort to AMMO – are these two linked in any way?

Max First, thank you, this was a nice talk and an interesting research effort. My question/idea: This may be an overkill, but if there is linking to professions, etc., maybe it can be connected to general dictionaries using, e.g. Ontolex-lemon instead/in addition to specific ontologies

Alessandro I was thinking the same in relation to the fact that some derivations are cross-cultural: for example, being related to priests is highly present in Greek (the prefix papa-)

Max Yes, exactly. It also could be interesting to compare cross-linguistically

Antoine (Only for reference, or if there's a lot of time for questions - it's not crucial!) I'm wondering if the authors have looked at how libraries handle and publish names, e.g. the Library of Congress Name Authority File (and the MADS format it uses). I guess this great work is much richer in detail, but maybe there's room for interoperability work in the future. On another topic, would it be possible to use DCAT for describing the dataset itself? This work is probably worth publishing on some portals that use DCAT.

Minna Thank you for the ideas, we will take a look at MADS and DCAT and indeed it might be a good idea and addition to describe the dataset.

Antoine Cool. Note that in fact MADS is not used a lot in a SW context, even though it has an RDF form. Maybe there are things in Bibframe, as an alternative..

Presentation: Fabian Hoppe, Tabea Tietz, Danilo Dessì, Nils Meyer, Mirjam Sprau, Mehwish Alam and Harald Sack. "The Challenges of German Archival Document Categorization on Insufficient Labeled Data".

Albert Curious about how open-ended the annotated categories are, and whether vocabularies were reused? Also, if you think the performance might be related to the training of word2vec with German Wikipedia?

Harald The categories have been provided by the archivists with respect to the available archival data (who organized them in a kind of hierarchical schema. It seems usual for German archivists, always to create new schemata depending on the current topic to be processed...). Unfortunately, labels can occur multiple times on different branches of this hierarchy. For your 2nd question, yes, the German wikipedia trained model is not the best for the task, since we are dealing with a historical subject and the language and topics used are 100 years old (and bound to a specific region in Germany). We plan to make use of historical newspaper archives for a better suited model.

Enrico This is very interesting! I am curious about how you join text and categories in the preparation phase for the embeddings. I would assume categories are somehow more important than the raw text - how did you combine the two, considering the algorithm wants a sequence of text as input?

Harald We try out different variants to come up with a representation for the categories ranging from simple category name embedding to aggregations of embeddings of members of a category as well as taking into account longer descriptions of the categories. However, experiments with labelled

data on classes with sufficiently available training data show that we can reach $> 75\%$ accuracy.

Coffee Break / 11:10 – 11:40

Presentation: Mattia Egloff, Alessandro Adamou and Davide Picca. "Enabling Ontology-Based Data Access to Project Gutenberg"

Albert "Undocumented RDFS classes and properties" Really curious about what those are, and which RDFS features they use? (domains, ranges, etc.) Also, what?s the contribution in inferred triples by the alignment layer? (assuming owl:sameAs?)

Rachele Just to point out that unfortunately Gutenberg site has been blocked by financial police...so sad

Alessandro (info on Italian closure here) the Italian court order contains an allegation that Gutenberg would be "illegally" profiting from their activity due to some unspecified ad revenue (which they don?t, unless you consider donations as such!)

Harald The same holds for Germany (more info on the blocking, cf `https://cand.pglaf.org/germany/index.html`)

Alessandro still accessible from Ireland, are we next? o_O

Enrico Accessible from the UK ? but so sad to hear that!

Albert Also accessible from the Netherlands, and indeed sad to hear this!

Presentation: Pejam Hassanzadeh, Eero Hyvönen, Esko Ikkala, Jouni Tuominen, Suzie Thomas, Anna Wessman and Ville Rohiola. "FindSampo Platform for Reporting and Studying Archaeological Finds Using Citizen Science"

Enrico Exciting project! I am curious about the policy for publishing people?s contributions and findings. The location of new archeological findings is quite a sensitive topic as there are countries that have a huge amount of heritage difficult to maintain or monitor, that can be subject to smuggling (e.g. Mexico, Italy). Any discussion on that in your project?

Eero Yes, lots of discussion. At the moment the feeling is that exact coordinate info will be published. It seems that professional archaeologists in Finland trust in amateurs more than in some other countries where only fuzzified data is published.

Albert Really cool project. I?m curious on the specific Linked Data features that were useful to users in the evaluation? E.g. reasoning, entity linking, etc.?

Eero In FindSampo Reporter, not much data linking is visible to the end-users. More important at the moment is integration of different systems such as GIS systems with the mobile system, and guiding the user to provide the data using harmonized terminology. In FindSampo Portal we are now focusing more on data linking, semantic faceted search, data analysis/visualization, and recommender systems.

Invited Keynote / 12.30

Chair: Enrico Daga (@enridaga)

Presentation: Antoine Isaac. "Europeana as a Linked Data (Quality) case"

Albert Europeana looks as impressive as usual, really nice :-) One question I had is if you do any sort of link rot/dereferencing maintenance? I understand you point items in Europeana to the original data provider's resources, but what happens if these change or become unavailable?

Antoine We're in trouble, so we try to encourage providers to be really careful about their changes. Note that we had (and are going to have again) a process that tries to catch some of these issues, by trying to recognize local identifiers in what is sent to us, and indeed do some de-referencing based on this. But we already know it won't catch everything

Albert A different question: sometimes I've found close matches to what I was looking for, but not exact ones. Is there a way I can request a specific item to the data providers?

Antoine I guess you would have to contact them, if it doesn't exist in what they've sent us. We're always eager to receive more material that fits user needs :-)

Albert A third one: On your massive vocabulary reuse, do you think the engineering of EDM adjusted well to standard ontology engineering practices? Or were there new/singular practices you had to implement to fit the domain?

Antoine excellent/tricky question, I hope I'm going to answer it right? In fact we have not followed the regular "formal" ontology engineering methods. It was not by pleasure (my first steps in SW were about such methodologies!). It's just that it would have been too hard to follow exactly a method and write down all the documentation. But in the end a lot of the best practices we've followed (and doc we've written) can be related to what is presented in these methodologies.

Jan Martin Which quality criteria did you use for source selection? And regarding vocabulary sources?

Antoine This was in one of my slides: Availability and access: open license, published as linked data Granularity, size and coverage: multilingual data, with a rather generic scope. But too generic or too large datasets can create too much ambiguity for the simple processes we have (e.g., enrichment) Quality: intrinsic aspects like correctness of representation Connectivity: good data sources are well-connected internally and externally to other datasets

Jan Martin How did you measure the "correctness of representation"?

Enrico On a similar angle, how do you deal with different and multiple (or even conflicting) perspectives on the same object/artwork? E.g. conflicting attribution statements (e.g. I am thinking about WikiData and their notion of truthy statements ?)

Antoine EDM uses a pattern from the OAI-ORE model whereby information from different sources is carried by different "proxies". We also re-use the Web Annotation models, which is a bit more intuitive way to represent annotations (but then it works better for individual data elements, not parts of graphs). We would have liked to use named graphs, but this required too advanced SW tools (our technical base is not an RDF quad store!)

Enrico Interested in the Linked Open Usable Data (LOUD) concept: what parts of Linked Data do you consider not usable? ... useful?

Antoine Very useful: URIs, links, lightweight data models that can be re-used, maybe pattern languages like ShEx/SHACL (though we couldn't use them yet). Not so useful and in fact often deceiving: OWL axioms and reasoning.

Break / 13:15 - 14:00

Session 2: Social History / 14.00 - 15.30

Chair: Alessandro Adamou

Presentation: Herminio García-González, Elena Albarrán-Fernández, Jose Emilio Labra Gayo and Miguel Calleja-Puerta. "Converting Asturian Notaries Public deeds to Linked Data using TEI and ShExML"

Alessandro Does the vocabulary used for diplomatic traits come as a subset of one that more generally deals with human psychological traits, dispositions etc.?

Alessandro You probably get this a lot, but it may come natural to anyone who is more familiar with SHACL than with ShEx, to ask what made you choose the other over the one, and how easy it would be for the constraints of your model to be ported to SHACL?

Albert Interesting use of schema.org, in CLARIAH we have lots of debate around schema.org modelling vs other vocabularies. What was important to you in your choice of schema.org terms?

Albert What do you think you could use as context for person disambiguation?

Enrico Very nice presentation and case study. How does ShExML compare, in terms of functionalities, with alternative RDF transformation tools such as RML or SPARQL GENERATE?

Presentation: Joe Raad, Rick Mourits, Auke Rijpma, Ruben Schalk, Richard Zijdeman and Albert Meroño-Peñuela. "Linking Dutch Civil Certificates"

Enrico That's an impressive resource! How far did you go in aligning people's identities. What are the major challenges in doing that?

Albert For this experiment in the Zeeland province we matched 270k-310k newborns in marriage certificates to brides/grooms, and 205k-244k parents of brides/grooms in marriage certificates to their own marriage certificate (depending on the Levenshtein distance). Scale is not really an issue due to our use of HDT and efficient data structures for computing Levenshtein distances. The major challenge is on the variability of person names, since people used to have many given names that sometimes changed among certificates.

Enrico Very interesting approach to entity linking, are you considering applying a similar strategy to other entity linking problems?

Albert Yes, definitely. The source code[1] is very generic, and we are working towards making all parameters dataset-agnostic. The idea is to have a dataset independent framework for entities in knowledge graphs that need to be linked using string similarity at large scale.

Enrico What about other entities, e.g. places - are they easily aligned to places of today or are there challenges in doing such linking?

Albert Interesting answer pointing at AMCO/gemeentegeschiedenis for temporal placenames, and HISCO for historical occupations

Discussion 1: The Semantic Web in Digital Humanities "Ecosystems"

Chair: Enrico Daga

- How many ecosystems?
  - Cultural Heritage, Research & Scholarship, Education, ...?
- The Knowledge Graph: does one-size fit all?

Alessandro Maybe it is not adding a fundamental difference?

Antoine The main difference is probably on the fact that KGs brings a more of private purpose to the original vision of LD & SW (and heterogeneity)

Albert In the past we envisioned a web of individuals while what is happening is a web of institutions. What we are seeing is an increasing role of institutions publishing and connecting knowledge graph

Albert Interesting article on authorities vs non-authorities `https://www.digitisednewspapers.net/2020-04-17-wrong-hierarchies/`

Enrico Is the distinction between types of ontologies (top-ontologies, domain ontologies) still valid? Or what alternative ways of characterising the modeling practices within the SW are useful, instead?

Antoine For example the notions of data patterns and best practices are very useful

Albert Maybe encapsulate 'sets' of semantic features needed to provide reasoning service for specific DH tasks – on ethical AI, whether a KG is respecting privacy or not, whether a dataset is biased or equally represents individuals, etc.. So guidelines for implementing these DH tasks through SW languages

[1] `https://github.com/CLARIAH/wp4-links`

- What is the role of Semantic Web technologies?
  - Are there specific aspects of DH research being enabled by SW technology?
  - What is the role of ontologies in DH ecosystems?
  - What is the role of reasoning in DH ecosystems?
  - How SW affects interoperability & distribution?
- What role Semantic Web research can have?
- Two extremes: open data / private data ? anything in between?
- Ontologies, Linked Data ... what next?

Coffee Break / 15:30 - 16:00

Session 3: Language / 16.00 - 16.50

Chair: Enrico Daga

Presentation. Rachele Sprugnoli, Francesco Mambrini, Giovanni Moretti and Marco Passarotti. "Towards the Modeling of Polarity in a Latin Knowledge Base"

Enrico Very interesting! A common notion in linguistics is that meaning is contextual. How - in your opinion - does this affect the quality or usability of a sentiment lexicon?

Rachele Given the results of our application to the Medea of Seneca, we think that the lexicon could be useful. But we sure need to improve the coverage.

Enrico How do you consider to evaluate the quality of the automatically generated silver GS?

Rachele We chose only derivational and semantic relations that were not ambiguous so to have a high-quality silver standard. For example, there are two in- prefixes but we used only the one expressing negation because the other can have different meanings. Details on the evaluation can be found in the LREC 2020 paper.

Enrico Latin changed heavily in the centuries/places, how do you (plan to) address the temporal or spatial variability in your project?

Enrico Are you planning to analyse and link to other latin resources?

Rachele We have a long list of resources to link. Examples are: the other Latin treebank and the Latin works of Dante Alighieri

Albert Really exciting work! I'm curious if you're planning some sort of "distant reading" evaluation? So e.g. visualizing how sentiment changes through a specific Latin text

Presentation. Tabea Tietz, Mehwish Alam, Harald Sack and Marieke van Erp. "Challenges of Knowledge Graph Evolution from an NLP Perspective"

Enrico Very interesting presentation! I particularly liked the case studies that capture the variety of aspects that related to KG evolution. I understand the paper is about the challenges ? but are you already thinking about strategies for making static KG incorporate ?some? of these dynamics?

Albert Really cool; I'd love to know more about further thoughts on using: (a) the typography (e.g. in your apfelstrudel example); and (b) semantic linking or language models to understand that the meaning of ?bomb? is very far away from what?s usual in recipe foods?

Discussion 2: WHiSe: feedback & community

Chair: Alessandro Adamou

- How can we better engage the community?
  - Online meetups
  - Webinars
  - Next editions of WHiSe

  Is the community happy with CEUR proceedings?
- Is the community happy with the venue (namely, Semantic Web conferences)? Shall we consider elsewhere?
- Are all areas of study concerning Humanities and Semantic Web well covered? Is the community missing any areas of study or specific DH topics
- Closing remarks and best paper announcement

END