IJDC | *Conference Pre-print*

Role of Content Analysis in Improving the Curation of Experimental Data

João Daniel Aguiar Castro
FEUP - INESC TEC, Portugal

Cristiana Landeira
FEUP

João Rocha da Silva
FEUP - INESC TEC, Portugal

Cristina Ribeiro
FEUP - INESC TEC, Portugal

Abstract

As researchers are increasingly seeking tools and specialized support to perform research data management activities, the collaboration with data curators can be fruitful. Yet, establishing a timely collaboration between researchers and data curators, grounded in sound communication, is often demanding. In this paper we propose manual content analysis as an approach to streamline the data curator workflow. With content analysis curators can obtain domain-specific concepts used to describe experimental configurations in scientific publications, to make it easier for researchers to understand the notion of metadata and for the development of metadata tools. We present three case studies from experimental domains, one related to sustainable chemistry, one to photovoltaic generation and another to nanoparticle synthesis. The curator started by performing content analysis in research publications, proceeded to create a metadata template based on the extracted concepts, and then interacted with researchers. The approach was validated by the researchers with a high rate of accepted concepts, 84 per cent. Researchers also provide feedback on how to improve some proposed descriptors. Content analysis has the potential to be a practical, proactive task, which can be extended to multiple experimental domains and bridge the communication gap between curators and researchers.

Submitted 15 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to João Daniel Aguiar Castro, INESC TEC, Faculty of Engineering, University of Porto, Porto, Portugal. Email: joaoaguiarcastr@gmail.com

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

As research policies lead institutions and researchers to adopt research data management (RDM) practices (European Commission, 2016), metadata activities are becoming embedded in research routines. One reward of investing in metadata production is that it favors data reuse (Thanos 2016), which may promote data citation and in turn lead to more data being deposited and reused (European Commission, 2016). As long as there are clear incentives and adequate tools, researchers are likely to engage more and more in RDM tasks (Pasquetto, Randles and Borgman, 2017). Researchers are domain experts and data producers, therefore they are well positioned to be key metadata producers as well (White, 2014), something they already do (Mayernik, 2011), particularly considering the generalized lack of expert staff (Wilson, 2007). On the other hand, scientific-oriented metadata is often supported by complex standards that researchers struggle to adopt (Qin and Li, 2013), if these are available at all.

At the University of Porto, under the TAIL project, we are focused in providing researchers with RDM tools to organize, describe and publish their data, and we are also incrementally improving our data curation workflow (Ribeiro et al., 2018). This workflow is mostly designed around activities in collaboration with researchers (Castro et al., 2015), such as informal meetings, interviews, data description and data publication. However simple it may seem, as we contact with more researchers, we realize that even basic concepts as metadata are hard to convey and that building a communication channel with the researchers is a key factor.

Some of our sessions with researchers, lasting at least one hour, have required a considerable amount of mental effort, without any satisfactory results. We consider that these unproductive interactions are a deterrent for a researcher with no previous engagement in RDM or that has to be convinced of its benefits. Thus, we increasingly started to explore content analysis (Stemler, 2001), in an ad-hoc fashion, with the goal of improving our workflow by 1) preparing data curators to talk with the researchers using domain terminology to ease communication; 2) making the data curators proactive in the definition of domain-specific metadata models. We believe that content analysis has great potential to improve the data curation workflow, and metadata production in particular, but the process needs to be formalized in order to become systematic.

In this paper we detail our content analysis approach to RDM in Section 3 and evaluate it with researchers dealing with experimental data in the sustainable chemistry, photovoltaic generation and nanoparticle synthesis domains in Section 4. We chose these domains for this proof of concept since experimental data is often reproducible if the procedures and relevant variables are well documented (Willis, Greenberg and White, 2012), while the diversity of research configurations calls for tailor-made metadata models rather than coverall standards.

Related Work

The value of metadata for RDM was reported in studies related to data sharing (Edwards et al. 2011) (Borgman, 2012), and on the improvement of data reuse (Greenberg and Fenstein, 2013), with an emphasis on the value of contextual metadata (Faniel and Yakel, 2011). Information about the methods to generate data is important contextual metadata, and it has been studied how different scientific metadata standards support the description of methods, however it was observed that there is potential for more comprehensive elements. In this context, research papers were identified as a rich source of information (Chao, 2014a)—however, to the best of our knowledge, content analysis is not a methodological technique usually associated to RDM, nor is it a common approach when developing tools for metadata production.

Nevertheless, an exploratory study based on the literature for soil science showed that journal publications hold relevant information for metadata production (Chao, 2014b). While

this study focused on the actual data description task rather than the selection of descriptors, it shows the need to systematize and the possibilities to extend the approach to other disciplines. Reading the papers that report on the experiments where the data were collected was also suggested as a task the data curator must perform, to quickly get a grasp of the research domain of the data being described, even if becoming a domain expert is not the goal (Wiljes and Cimiano, 2012).

Since curators in institutional data services are expected to describe many data sets from a myriad of domains and in a very limited time, one must look towards automating the process as much as possible. Automated methods such as Named Entity Recognition, present in packages such as CoreNLP (Finkel, Grenager and Manning, 2005), or Keyword Extraction, implemented in the YAKE! framework (Campos et al., 2018) can be used to highlight the most important concepts referred in the research texts that usually are related to datasets. At the same time, they can help highlight relevant parts of a document so that the curator can more easily spot possible metadata to include in the dataset record.

With this in mind, we have recently applied content analysis to assist in metadata production in specific domains. Our goal was to discover those concepts that could be mapped to domain-specific descriptors, in our case properties from different ontologies, which were drawn from DBpedia or the Linked Open Vocabularies (LOV) catalog (Vandenbussche, Atemezing and Vatant, 2014) after a keyword extraction step using both CoreNLP and YAKE!. The results showed the complexity of the task, as even after keyword extraction there is a large set of possible ontology properties to choose from, and highlighted the indispensable role of the curator in the process for systematically validating and complementing the results of any automatic tool (Monteiro, Lopes and da Silva, 2018).

Information extraction from documents has been applied to RDM before: in the chemistry domain, for example, it has been used for the development of ontologies and predictable models from data. The result was considered useful to deal with significant amounts of data and structured documents, but not effective when applied to less structured descriptions of chemical procedures (Townsend et al., 2004). Accordingly, chemistry librarians have argued that humans are able to efficiently summarize and to present information as opposed to the limitations that a fully automated approach might entail, such as many false positives in the selected concepts and overlooked details (McEwen and Li, 2014).

Methodological Approach

During our collaboration with researchers at the University of Porto, we had the opportunity to develop metadata models for several domains (Castro et al., 2015). In experimental domains we relied solely on an interview form, complemented with descriptors that researchers were able to suggest based on their perception. These were often influenced by the difficulties of reaching an agreement on metadata conceptualization.

Our approach to content analysis in the data curation workflow comprehends the identification of relevant segments of text in scientific publications, in a particular section reporting the experimental set-up, like the methodological approach or the experimental configuration. The experimental set-up section is particularly interesting as it systematically describes the parameters of a given experiment, that is, what precedes and provides the context for the production of data, therefore a requirement for scientific metadata (Qin, Ball and Greenberg, 2012). On the other hand, sections covering the results, although important to know more about the domain, are the output of an experiment with a greater focus on the data itself than on the context of production. Moreover, if the proposed approach entails the integral reading of the papers it would be a counterproductive task. However, a brief reading of the introductory section of each publication, or any other additional section, will give the curator a broader scope of the domain that may be useful to the overall task and to the conversations with the domain experts that follow.

The selected text segments are those where the researchers assigned a specific value to a property or made an environmental characterization. For instance, if the researcher writes "The ozonation and the experiments with ozone-based AOPs were conducted in a bubble-column semi-batch reactor" we infer that the ozonation reactor is a candidate metadata element.

To ensure that the process is as realistic as possible from the data curator standpoint without being dependent on their degree of specialization, we assigned the content analysis task to a curator with limited RDM expertise, and a time frame of no more than two weeks, shared with other tasks. We also selected a small corpus of publications for each domain, considering that content analysis in a large sample might be more appropriate for an automatic approach or if the goal is to retrieve more values for the development of controlled vocabularies. Also, we assume that if some piece of information is relevant in a particular domain that kind of information would be present even in a small number of papers.

After processing the text and setting up a list of metadata candidates, we have prepared an informal metadata template in a shared document for the researchers to fill in. This template is a simple two-column form, with the proposed descriptors in one column and an empty one in the other for the researcher to add the corresponding value. We then asked researchers to insert values for the descriptors they considered appropriate or else to comment on how to make the descriptor more appropriate, in case they believed that the concept could be improved. If the researchers did not insert a value we requested a further comment on the reason for this. The experimental domains to use in this experiment were determined by existing contacts with researchers producing data and on their agreement to participate in an evaluation session. Hence, we applied the approach in three experimental domains. The publication corpus for each domain was collected based on keywords related to the experimental configurations that the researchers in the three cases perform regularly, or upon recommendation by the researcher of a paper that describe similar experimental work. We have worked with three publications for each domain. The dataset, with reference to the publications subject to content analysis, the corresponding text segments extracted and the proposed descriptors is available online - the proposed descriptors were specified in Portuguese (Castro and Landeira, 2019). One researcher from each domain participated in the evaluation.

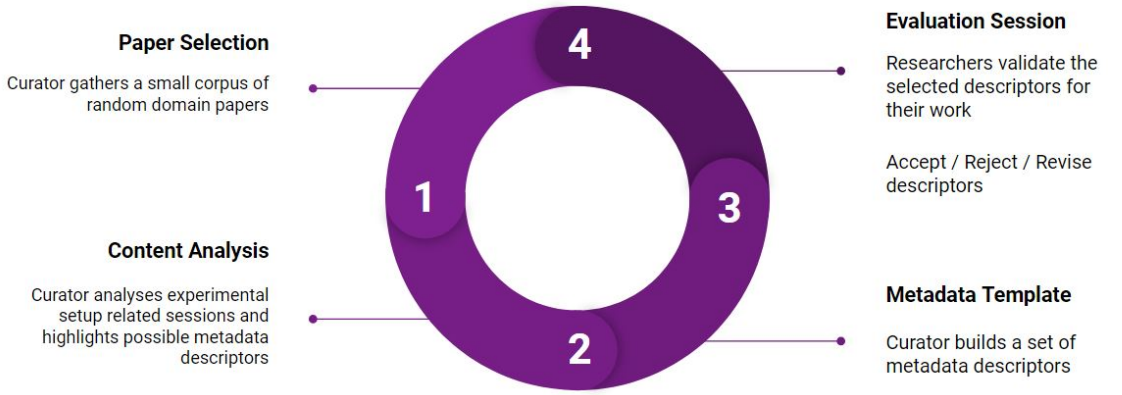

Figure 1: Methodological approach; from content analysis to the metadata template evaluation

Sustainable Chemistry, degradation of pollutant particles

Human activity waste accumulates in the environment and contaminates water, soil and atmosphere, triggering all sorts of hazards. Global warming, water shortage, health risks and malformation are some of the issues amplified by pollution. It is therefore important to eliminate pollutants or make them less offensive to the environment, and solutions to achieve these goals are being studied.

To reconstruct the context of an experiment related to the degradation of pollutant particles, it is necessary to capture the properties of multiple samples being studied, to record the instruments applied in the experimental workflow, their characteristics and calibrations, according to their influence in the final results. Metadata for the methods and techniques applied give a broad view of the experiments. On top of that, details of to the duration of certain experimental events, measurements, and environmental controlled conditions, contribute to metadata quality and to the trustworthiness of data.

Photovoltaic Generation, thin film experiments

Solar energy is growing as a renewable and clean energy type. Photovoltaic energy generation is a method to convert solar light into electric energy. This transformation happens through semiconductors that can release energy when stimulated by light. Due to the demand for clean energy solutions, ensuring the provision of sustainable electrical power, many studies, as the study of thin films and the study of optical properties of copper, gallium, selenium and others, have been developed.

This kind of experiment involves different methods and techniques that influence the experimental configuration. For instance the effect of the so-called annealing temperature is only required if the researcher is adopting a technique that submits the sample to this factor. For the contextualization of photovoltaic thin film data one needs to know the technique of elaboration of the absorbing layer, the precursors used for the elaboration of the thin films, the optimized experimental conditions, the deposit substrate and its properties, the electrical and optical properties of the thin films produced. The researcher conducting thin film experiments was not available in person, so the evaluation task was done remotely. We shared the link to the metadata template file for this domain, and used an online chat platform to provide instructions to the researcher and get her feedback.

Nanoparticle Synthesis

Nanoparticle synthesis refers to the methods for creating nanoparticles, and the development of this type of experiment is relevant since there is a wide range of applications in a diversity of areas. Research opportunities in this domain are rich given the capacity to revolutionize the characteristics and functionalities of materials on a nano scale.

In construction it can have the objective to contribute to improve building conditions with materials that last longer, or with additional functionality. Nanoparticles can be applied, for instance, to enable the materials to autonomously remain clean. Moreover, this line of research has an impact in health applications, in diagnosis, transplants and tissue engineering, as nanostructures allow cell interactions previously prevented due to their size.

For this case a total of 74 potential descriptors were identified by the curator. However, after the two previous evaluations, we choose to represent in the template only a smaller set of descriptors to see if the researcher, with a less exhausting task ahead, since the number of descriptors to evaluate was significantly smaller, would show a different behaviour in the task. For example, more time considering options, not being influenced by a large number of concepts to have more room to suggest new ones and to verify if these were in line with those that we had omitted. Therefore, the number of descriptors in the metadata template for this evaluation was only 23.

Metadata Template Evaluation

For each case we prepared a form with the proposed descriptors, where some are common to the three templates. We also provided the researchers with instructions to fill in or to make a comment when the insertion of a value was not considered.

Table 1 depicts the overall results. We presented a total of 139 descriptors, from which 117 were accepted by the researchers, some with suggested revisions. The remaining 22 were rejected, most viewed as unnecessary by the photovoltaic generation researcher, while some were repeated concepts with overlapping semantics or were not understood. The overall acceptance ratio was 84 per cent.

Whenever a proposed descriptor is rejected, this does not mean that it is not suitable for the specific domain. It may well happen that another researcher in the same domain applies some techniques with properties that are unfamiliar to the researchers in our case. The same is true for potential descriptors not identified by the data curator, as different subareas may demand additional ones.

Table 1. Overall results

Descriptor evaluation	**Sustainable Chemistry**	**Photovoltaic Generation**	**Nanoparticle Synthesis**
Directly accepted	38	42	18
Revised	15 descriptors – e.g. Interfacial area- Interfacial distance	3 descriptors– e.g. Optical transmittance - Transmittance	3 descriptors – e.g. Laser pulse width – Laser pulse
Suggested	Solution concentration	Dielectric constant real/ imaginary part	-
Not needed	Catalyst wavelength	11 – e.g. Radio frequency	Reducing agent
Repeated	3 descriptors – e.g. Chemical demand (oxygen) = (Ozone) mass flow rate	-	-
Not understood	3 descriptors – e.g. Spectral measurement instrument	-	Passivation molecule concentration

Sustainable Chemistry evaluation

In the sustainable chemistry metadata template we included a total of 60 metadata fields, from which 53 were understood by the researcher. From these, 38 were directly filled in or approved, while the researcher has suggested improvements on the remaining. On the other hand, 8 descriptors were rejected; either perceived as redundant, such as the *Gas superficial velocity*, understood but not used or did not make sense for the researcher (see Appendix, table 2). The evaluation was concluded in a single session taking about one hour. At the end we had an informal conversation with the researcher to obtain additional feedback.

The researcher stated that there is a need to record the experiments in laboratory notebooks that work as a minute of experimental configurations, especially when unexpected events occur. When asked if the metadata template was comprehensive enough to describe data on the degradation of pollutant particles, and capture the minute information, the answer was positive, but the researcher also noted that more elements are required. Nevertheless, the researcher pointed out that some fields should be more specific, e.g. the Solution might be replaced by Solution concentration, from which the solution can be identified. Another suggestion was the Analysis method, rather than Polyphenol analysis method, which was considered, over-specialized.

The metadata template for this case included both a generic Instrument field, and additional ten descriptors for the identification of specific instruments, e.g. Light radiation measurement. In this case the researcher rejected Instrument, given the difficulty posed by the diversity and number of instruments used in a single experiment. Fill in the generic Instrument descriptor with information about all the specific types of instruments that were used in an experiment and their purpose was perceived as a burdensome task by the researcher. Therefore, the researcher prefers to have available descriptors for each specific instrument and only introduce a value, for instance, its model. From the list of specific instrument descriptors proposed only the Spectral measurements instrument was rejected.

To make the data curator workflow more efficient the researcher suggested that the process might include metadata generated by the many instruments that comprise the experimental set up. However, it was recognized that it would be a challenge for the data curator to gain access to all instruments, as they are spread across different laboratories and require authorization by the lab coordinators. This is something hard even for the researcher, as noted.

Photovoltaic Generation evaluation

Since the evaluation on the photovoltaic generation case was done remotely, the metadata template was completed in more than one interaction, over a few weeks, according to the agenda of the researcher. Although this process was constrained by the lack of personal feedback, it also offered the possibility to obtain results with less assistance from the data curator.

In this case we list a total of 56 fields from which 45 were understood, with the researcher proposing improvements for two of them. The 11 remaining were rejected with the justification that the metadata element was useless, such as the Electrical resistance and the Spraying time descriptors, as listed in Appendix, table 3.

The researcher confirmed the proximity between the proposed concepts and the information they usually record in a notebook or text file, but alluded to the lack of a metadata element for the precursor's concentration, which was not included in the metadata template. Thus, more metadata fields are required to capture all the relevant information in this case.

Similarly to the sustainable chemistry case, the photovoltaic energy researcher made considerations about the granularity of concepts. For the Optical transmittance descriptor the researcher suggested the adoption of the broader term Transmittance, while for some other cases more specific descriptors were suggested, such as subdividing Dielectric constant into Real part of dielectric constant and Imaginary part of dielectric constant. The description of instruments is also relevant, however only the Instrument field was available in the template, given the lack of expertise of the data curator to assign different types of instruments in this context. Nevertheless, this field was completed without further comment by the researcher.

Nanoparticle synthesis evaluation

As for the nanoparticle synthesis evaluation, from the 23 descriptors in the template a total of 20 descriptors were accepted, 16 directly filled in (Appendix, table 4). The researcher recommended improvements on three others, while one was considered not very precise. The remaining three were rejected, considered meaningless or not needed for the researcher. The metadata template also included three descriptors for the representation of very specific types of instruments, the *Optical properties analysis instrument*, the *Sample synthesis instrument* and the *Radiation emission instrument*. The researcher added a description in two of them without further comments, although it has shown preference for the specification of the instrument when asked a generic *Instrument* descriptor would be suitable. A field for the description of the instrument producer would also be useful in some experimental contexts, according to the researcher.

For the descriptor Sample Coat the researcher stated that this is not the most suitable term, but recognized that many colleagues can use it regularly, not knowing what the alternative term might be. The proposed *Reducing agent* descriptor was one of the concepts that was understood as useful but that was not used in the experiments performed by the researcher in this evaluation.

For the descriptor *Laser pulse width* the introduced value was *248 nm (wave-length); 500 mJ (pulse energy); 10Hz (pulse frequency); 20 nS (pulse duration*). This suggests that the researcher would prefer a structured description instead of having to fit this information in an unstructured descriptor. It is also important to highlight that the list of hidden descriptors from the metadata template already included most of the necessary fields for the needed representation, namely *Pulse wave-length*, *Pulse frequency* and *Pulse duration*, while the *Laser energy per pulse* was a potential descriptor identified by the curator. However, if the *Radiation emission instrument* is available there is no need to record the wave-length, according to the researcher. When asked if there were missing descriptors, the researcher said that those presented were enough and that the metadata would be useful for other researchers as well.

After the evaluation of the 23 proposed descriptors, taking into account the duration of the session (about 30 minutes) and the availability demonstrated by the researcher, we suggested a quick observation of the remaining descriptors. From the list of 51 remaining descriptors, one was seen as ambiguous. At a certain point of the evaluation the researcher concluded about the importance of the descriptors that *"their use will depend on the experiment"*. It depends of the method, technique, sample and instruments chosen, for instance the *Synthesizing vessel dimension* is only necessary if a synthesizing vessel is used.

When asked if the session was useful and if it was easy to participate in the task, the answer was positive, yet the researcher consider that if the descriptors in the template were organized and not *"all mixed"* it would ease the description, acknowledging that a correct organization of concepts would be a difficult task for someone who is not an expert in the scientific domain.

The nanoparticle researcher also mentioned that there is a need to annotate the experimental context and that she *"cannot work in chaos"*, and prior to this experiment already discipline herself to annotate all the contextual information in her experiments. These annotations are made using slides, so a presentation is always ready when necessary. Other methods, like keeping a notebook, were explored but the researcher could not organize the information so efficiently.

Discussion

The assumption in this work was that content analysis positively impacts the data curator workflow by improving the communication with the researchers and making the data curator proactive in the definition of domain-specific metadata models. With the metadata template evaluation we obtained tangible indicators that support this hypothesis. Our past experience carrying out RDM tasks with researchers makes it possible to reflect on more elusive indicators.

The content-analysis step is likely to decrease the communication gap between researchers and data curators. This is due to the increased awareness and interest on the researcher side, and also to the domain expertise gained by the data curator.

In our experiments with a large sample of groups, we systematically request feedback from researchers on how to make the interaction better, and many consider the adoption of domain terminology as something that helps them to quickly understand RDM benefits and practices.

Performing manual content analysis provides domain knowledge to the data curator even before the first interaction and throughout the process, facilitating communication with the researcher. Hence, in the interaction with the researchers, the data curators can adopt domain concepts to illustrate RDM scenarios, as opposed to more generic metaphors, e.g. based on Dublin Core metadata. Disciplinary examples are something that researchers tend to ask for.

By showing researchers a template with familiar concepts we demonstrate interest in their domains, establishing a productive, empathetic relationship that makes talking about metadata less demanding. Moreover, starting an interaction with good communication also leads to faster input from the researchers and raises their awareness in an effective way. For instance, it has motivated the sustainable chemistry researcher to make a suggestion as soon as the metadata template evaluation was completed.

Additionally, the high acceptance rate of the descriptors in the metadata template evaluation provides evidence of the advantages of performing content analysis before the first meeting with the researchers. In this line, a data curator can assume with some confidence that many potential descriptors resulting from content analysis will be included in a domain-specific metadata model. Moreover, the identification of descriptors was recognized as a realistic activity from the data curator point of view, since it was performed in a reasonable time frame and did not required in-depth domain expertise.

The evaluation with researchers shows that the definition of metadata models must counterbalance generic and specific descriptors. Highly-specific descriptors were filled by the researchers very quickly, while the absence of some specific descriptor apparently limits the metadata, as shown in the sustainable chemistry case with the descriptor for the type of instrument. The description provided by the researcher in the nanoparticle case, for the Laser pulse width, showed that a greater specification of descriptors can make metadata production smoother. On the other hand a high number of descriptors entail more time spent in the selection of descriptors, which can be perceived as a barrier to the adoption of very specialized RDM tools.

Another possible limitation is that the properties being captured can change significantly from one sub-domain to the other, requiring the use of additional descriptors for each case, if we aim for metadata requirements at such level. Withal, a higher degree of specificity is desirable when metadata tools or platforms allow the combination of metadata elements with respect to the diversity of experiments, techniques and datasets, a given researcher may need to describe. A scenario where researchers do not find a required descriptor, or only have generic descriptors available and are unable to record all the relevant information, should be avoided.

In this evaluation the data curator adopted an exhaustive content analysis approach and still the task was seen as practical. We believe that if the data curator adopted a principle of minimal effort, only capturing high-level descriptors that might still be enough to start the conversation with the researchers and let them contribute with finer metadata requirements. Likewise, the greater the number of descriptors specified by the curator the narrower the possibility will be of the researchers contributing descriptors of their own, hence making it harder for the curator to determine which the most relevant descriptors for each domain are.

Regardless of its merits there is still room to improve this approach. An example is the introduction of tools for entity extraction from the texts provided by researchers at the start of the process. While these automated approaches show great potential in helping the curator navigate larger collections of texts, the results of our past work with keyword extraction approaches for metadata production show that they cannot be seen as a replacement for the expertise and engagement that the curator brings to the process, but rather as a complement.

Conclusion

Following this evaluation we have ongoing work to verify if similar outcomes can be obtained in other experimental domains. So far, as we diversify the domains, we have found that the data curator can 1) simply reuse or improve concepts, since experimental data share many properties; 2) develop skills that reduce the effort when addressing new domains.

We regard content analysis as a complementary task in the development of metadata tools, such as domain-specific ontologies, that we have evaluated with researchers in data description sessions using Dendro, a data organization platform aimed at data organisation and description early in the research workflow (da Silva, Ribeiro and Lopes, 2018). However, even a very specific domain or a particular type of experiment can encompass several techniques, each with its own metadata requirements. Our expectation is that as the number of descriptors grows researchers can then combine suitable descriptors for each dataset, depending on the experimental setup that originate them.

Automatic ontology-learning approaches are also being considered under the TAIL project (Monteiro, Lopes and da Silva, 2018). The results are promising, especially for recall, but precision needs to be improved. From a data curator perspective, manual and automatic approaches can go hand in hand.

Although automatic content analysis can expedite the process and deal with larger corpora, making sense of the large number of automatically extracted concepts still requires decision making, considering the subjectiveness involved in giving context to the extracted concepts, in order to infer the descriptors.

To conclude we do not anticipate manual content analysis as an activity to be performed regularly by a data curator. In fact, many researchers already have well-detailed experimental protocols and scientific metadata standards are available, some in experimental domains (Willis2012). Even so, this approach can be adopted as long as there is a need to define metadata requirements from the beginning or to specialize extant tools.

Acknowledgements

This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within project TAIL, POCI-01-0145-FEDER-016736. João Aguiar Castro is supported by research grant PD/BD/114143/2015, provided by the FCT - Fundação para a Ciência e a Tecnologia.

References

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63 (6). doi.org/10.1002/asi22634

Campos, R., Mangaravite, V., Pasquali, A., Alípio, M. J., Nunes, C. & Jatowt, A. (2018). A Text Feature Based Automatic Keyword Extraction Method for Single Documents. Proceedings of the 40th European Conference on Information Retrieval. Retrieved from https://repositorio.inesctec.pt/bitstream/123456789/7622/1/P-00N-NF3.pdf

Castro, J. C., Amorim, R., Gattelli, R., Karimova, Y., da Silva, J. R. & Ribeiro, C. (2017). Involving Data Creators in an Ontology-Based Design Process for Metadata Models. In Developing Metadata Application Profiles. IGI Global.

Castro, J. C., Perrotta, D., Amorim, R., da Silva, J. R. & Ribeiro, C. (2015). Ontologies for research data description: a desing process applied to Vehicle Simulation. Metadata and Semantics Research Conference

Castro, J. A., & Landeira, C. (2019) Content Analysis of Publications in Experimental domains. INESC TEC research data repository. https://doi.org/10.25747/kh8b-xx50

Chao, T. C. (2014a). Enhancing metadata for research methods in data curation. Proceedings of the American Society for Information Science and Technology, 51 (1) doi.org/10.1002/meet.2014.14505101103

Chao, T. C. (2014b). Identifying Description Indicators for Research Data from Scientific Journal Publications. iConference Proceedings. doi:10.9776/14366

Greenberg, J. & Feinstein, E. M. (2013). Metadata Capital in a Data Repos-itory. Metadata Reuse: A Growth Indicator. International Conference on Dublin Core and Metadata Applications.

Edwards, P. N., Mayernik, M., Batcheller, A., Bowker, G. & Borgman, C. L. (2011). Science friction: Data, metadata and collaboration. Social Studies of Science, 41 (5). doi.org/10.1177/0306312711413314

European Commmission (2016). Guidelines on FAIR Data Management in Horizon 2020. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission (2018). Turning FAIR into reality. doi:10.2777/1524

Finkel, J. R., Grenager, T. & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics Workshop on the Semantic Publishing. doi:10.3115/1219840.1219885

Mayernik, M. (2011). Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators. Retrieved from: SSRN https://ssrn.com/abstract=2042653. doi.org/10.2139/ssrn.2042653

McEwen, L. & Li, Y. (2014). Academic librarians at play in the field of chemistry informatics: Building the case for chemistry research data management. Journal of Computer-Aided Molecular Design, 28 (10). doi:10.1007/s10822-014-9777-4

Monteiro, C., Lopes, C.T. & Rocha da Silva, J. (2018). Supporting Description of Research Data: Evaluation and Comparison of Term and Concept Extraction Approaches. In Digital Libraries for Open Knowledge. doi:10.1007/978-3-030-00066-0_44

Pasquetto, I. V., Randles, B.M. & Borgman, C.L. (2017). On the Reuse of Scientific Data. Data Science Journal, 16. doi.org/10.5334/dsj-2017-008

Vandenbussche, P., Atemezing, G. A., Vatant, B. (2014). Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web. Semantic Web Journal, 1. doi:10.1002/asi.22683

Qin, J. Ball, A. & Greenberg, J. (2013). Functional and ArchitecturalRequirements for Metadata: Supporting Discovery and Management of Scientific Data. International Conference on Dublin Core and Metadata Applications

Qin, J. & Li, K. (2013). How Portable Are the Metadata Standards for ScientificData? A Proposal for a Metadata Infrastructure. International Conference on Dublin Core and Metadata Applications

Ribeiro, C., da Silva, J. R. Castro, J. A., Amorim, R. C., Lopes, J. C. & David, G. (2018). Research Data Management Tools and Workflows: Experimental Work at the University of Porto. IASSIST Quarterly, 42 (2). doi.org/10.29173/iq925

da Silva, J. R., Ribeiro, C. & Lopes, J. C. (2018). Ranking Dublin Core descriptor listsfrom user interactions: A case study with Dublin Core Terms using the Dendro platform. International Journal on Digital Libraries, 20. doi.org/10.1007/S00799-018-0238

Stemler, S. (2001). An Overview of Content Analysis. Practical assessment,research & evaluation, 7 (17)

Thanos. C. (2016). Scientific Data Reusability: Concepts, Impediments and Enabling Technologies. Publications, 5 (1). doi:10.3390/publications5010002

Townsend, J., Adams, S. E., Waudby, C., de Souza, V. K., Goodman, J. M. & Murray-Rust, P. (2004). Chemical documents: machine understanding and automated information extraction. Organic & biomolecular chemistry 2, (22). doi:10.1039/b411033a

White, H. C. (2014). Descriptive Metadata for Scientific Data Repositories: AComparison of Information Scientist and Scientist Organizing Behaviors. Journal of Library Metadata, 14 (1). doi.org/10.1080/19386389.2014.891896

[journal article] Willis, C., Greenberg, J. & White, H. (2012). Analysis and Synthesis of Metadata Goals for Scientific Data. Journal of the Association for Information Science and Technology, 63 (8). doi:10.1002/asi.22683

Wiljes, C. & Cimiano, P. (2012). Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology. Proceedings of the Workshop on the Semantic Publishing. Retrieved from http://ceur-ws.org/Vol-903/paper-07.pdf

Wilson, A. J. (2007). Toward Releasing the Metadata Bottleneck. A Baseline Evaluation of Contributors-supllied Metadata. Library Resources & Technical Services, 51 (1)

Appendix

Table 2. Results from the Sustainable Chemistry template evaluation

Curator input	Researcher comment / recommendation
Chemical compound	It causes doubt; everything is a chemical compound, so it has no specificity.
Mass transfer coefficient	Do not know what to describe.
Oxidation agent potential	The value would be the same for Oxidation potential.
Interfacial area	Not all samples have an interfacial area.
Chemical demand (oxygen)	It does not make sense this way. The two are alike (ozone mass flow rate).
(Ozone) mass flow rate	Gas phase flow.
(Pollutant) Ph	Solution Ph
Polyphenol	Do not understand. Ask for the formula, and for its quantity ask mass or volume.
Molecular mass	It is necessary to define the molecular mass of what?
Impurity tenor	Very specific, by knowing the purity degree you calculate the impurity tenor.
Aqueous solution	The aqueous repeats the solution idea. Is necessary to define the type of solution. Cleaning solution / Acid solution.
Pollutant particle size	Particle size
Remaining accepted descriptors	Atmosphere conditions; Total carbon; Total organic carbon; Reagent; Oxidant agent; Chemical element; Control solution; Photocatalytic activity; Solar light intensity; Sample crystallite size; Sample pore volume; Sample reference; Sample centrifuged amount; Sample drying temperature; Adsorbent area; Adsorbent ash tenor; Adsorbent particle size; Adsorbent molecular formula; Particle removal technique); Surface area measurement technique; Electromagnetic radiation measurement Instrument; Ozonisation reactor instrument; Ph measurement instrument; Absorbance measurement instrument; Light radiation instrument; Light intensity measurement instrument; Surface area measurements instrument; Catalysts analysis instrument); Photocatalytic reaction vessel instrument; Ozonation time, Light intensity measurement time; Suspension stirring time
Inorganic carbon	Do not use the concept but I recognize the concept.
Gas superficial velocity	The same as gas phase flow.
Remaining Rejected Descriptors	Ozone partial pressure (gas phase); Ozone interfacial concentration; Catalyst wavelength; Spectral measurements instrument; Sample diluted centrifuged amount; Absorbance

Table 3. Results from the Photovoltaic Generation template evaluation

Curator input	Researcher comment / recommendation
Optical transmittance	Transmittance
Dielectric constant	Dielectric constant *real part / *imaginary part
Absorbent layer production technique	Absorbent layer manufacturing technique
Remaining accepted descriptors	Method; Chemical compound; Band gap; Deposition potential; Semiconductor type; Potential rage; Complexing agent; Reaction type; Bath configuration; Characterization technique; Deposition time; Gap energy; Refractive index; Extinction coefficient; Compound yield; Compound absorption coefficient; Compound physical state; Sample drying; Sample drying temperature; Sample drying time; Substrate type; Substrate dimension; Substrate cleaning method; Substrate temperature; Working electrode; Electrode reference; Electrode counter); Annealing time; Annealing temperature
Electrical resistance	I cannot find the exact resistance.
Spraying time	It is for a technique I do not use.
Remaining Rejected Descriptors	Compound viscosity; Compound boiling point; Reagent; Solution matrix; Photon energy; Cathodic sputtering source; Radio frequency; Sample power; Temperature stabilization time

Table 4. Results from the Nanoparticle Synthesis template evaluation

Curator input	Researcher comment / recommendation
Sample coat	Not the most suitable term.
Sample concentration	Sample mass
Laser pulse width	Pulse energy
Sample heating time	Deposition time
Remaining accepted descriptors	Sample; Sample producer; Sample coat dimensions; Stabilizer; Particle size; Solution; Instrument; Optical properties analysis instrument; Sample synthesis instrument; Radiation emission instrument; Pulse duration time; Synthesis temperature; Synthesis method; Characterization technique; Atmosphere conditions; Substrate
Passivation molecule concentration	Does not make sense.
Reducing agent	I understand but I do not use it in my experiments.
Remaining Rejected Descriptors	Milli-Q resistance