Storing combustion data experiments: new requirements emerging from a first prototype
Position paper

Gabriele Scalia[1], Matteo Pelucchi[2], Alessandro Stagni[2], Tiziano Faravelli[2], and Barbara Pernici[1]

[1] Department of Electronics, Information and Bioengineering
[2] Department of Chemistry, Materials, and Chemical Engineering Giulio Natta
Politecnico di Milano,
[name.lastname]@polimi.it
http://www.polimi.it

Abstract. Repositories for scientific and scholarly data are valuable resources to share, search, and reuse data by the community. Such repositories are essential in data-driven research based on experimental data. In this paper we focus on the case of combustion kinetic modeling, where the goal is to design models typically validated by means of comparisons with a large number of experiments.

In this paper, we discuss new requirements emerging from the analysis of an existing data collection prototype and its associated services. New requirements, elaborated in the paper, include the acquisition of new experiments, the automatic discovery of new sources, semantic exploration of information and multi-source integration, the selection of data for model validation.

These new requirements set the need for a new representation of scientific data and associated metadata. This paper describes the scenario, the requirements and outlines an initial architecture to support them.

Keywords: experimental data, explorative approaches, combustion modeling

1 Introduction

The collection of experimental data for scientific research is becoming more and more important for the validation of the research results. One of the current goals is also the ability and effectiveness of sharing experimental data among researchers, in order to increase their quality and reproducibility and to derive and validate new research results.

In particular, in this paper we focus on experimental data in the combustion domain, and in particular in chemical kinetics, where a number of initiatives have been developed to collect experimental data in a systematic way, to be shared in the research community. Recent developments of experimental data repositories for chemical kinetics (e.g., ReSpecTh [4], CloudFlame [2], ChemKED [1], PrIMe

[3]) aim to collect and store the increasing number of basic and complex experimental measurements of reacting and non-reacting combustion phenomena (or properties) in more efficient machine-readable formats (e.g. XML, YAML etc.). In parallel, EU-funded projects are pursuing the challenging goal of defining community data reporting standards[3] to overcome instances of incomplete, inaccurate, or ambiguous descriptions of fundamental data, both in the past and in the recent scientific literature.

The urgent need of improving the infrastructure supporting the reuse of scholarly data has been highlighted in literature. To facilitate this effort, general guidelines have been proposed for creating and managing repositories, like the FAIRness [30] (being findable, accessible, interoperable and reusable) or the "pyramid" of needs for data management that span from being simply *saved* to being *shared* to ultimately being *trusted* [12].

On top of these needs for data, new analytics requirements arise. The new requirements are mainly related to the *semantics of data*. Taking into consideration the semantics of the stored data, necessary to improve tasks like acquisition, exploration and validation, brings new challenges. For example, "this necessitates machines to be capable of autonomously and appropriately acting when faced with the wide range of types, formats, and access-mechanisms/protocols that will be encountered during their self-guided exploration" [30].

If a scientific repository with its basic tools for import/export can improve the efficiency of many tasks, like structurally retrieving certain experiment types, on the other side taking into account the semantics allow improving the *effectiveness* of data management. For example, many different functionalities that can exploit the scientific information conveyed by data and evaluate it not only for the quality of the available data/information itself, but with respect to the available models for that particular scientific experiment described by those data.

Others emerging requirements are arising in terms of the ability of retrieving data in an exploratory way [13, 28, 14]. This involves searching non obvious relations among data exploring possible research directions, and being able to assess the quality of the retrieved information.

The goal of the present paper is to discuss on new requirements emerging from the development and use of existing repositories of experimental data in the domain of kinetic modeling of chemical processes such as combustion [21, 27]. While these new requirements are investigated in this specific domain, they are general and may be extended to many other fields in the wider domain of scientific experimentation.

The paper is structured as follows. Section 2 introduces the domain presenting the scenario, Section 3 presents the emerging requirements and Section 4 sketches an architecture to support them.

[3] http://www.smartcats.eu/wg4/task-force/

2 Scenario

Combustion kinetic modelling has been driving the development of more efficient fuels and combustion technologies (e.g. internal combustion engines, gas turbines, industrial furnaces etc.) for the last 30 years. As a matter of fact, chemical kinetics determines the reactivity of a given fuel or a fuel mixture; thus, a better understanding of the effects of a specific chemical compound on combustion performances and emissions allows the tailoring of a fuel or a fuel blend for an existing infrastructure or vice versa [5].

The Chemical Reaction Engineering and Chemical Kinetics (CRECK) research group at Politecnico di Milano deals on a daily basis with the development and update of such kinetic models.

The development and update of reliable kinetic models is a rather challenging task, directly reflecting the intrinsic complexity of combustion phenomena, and is one of the fields of research of the CRECK modeling group[4]. Such models typically involve $\sim 10^2 - 10^3$ chemical species connected by a network of $\sim 10^3 - 10^4$ elementary reaction steps. Moreover, a combustion kinetic model hierarchically develops from small chemical species (e.g. hydrogen, methane, etc.) up to heavier compounds typically found in commercial fuels (gasoline, diesel and jet fuels). For this reason, any modification in the core mechanism significantly propagates its effects to heavier species making continuous revisions and updates mandatory to preserve the reliability of the model.

From an operational perspective, the iterative validation of such models (Figure 1) strongly relies on extensive comparisons of results from numerical simulations with an enormous number of experimental data covering conditions of interest for real combustion devices. The key step in such procedure consists in the objective and automatic assessment of model performances, properly taking into account experimental uncertainties, and avoiding time consuming and unsustainable qualitative comparisons. Analysis tools (e.g., sensitivity analysis) allow highlighting relevant model parameters and drive their refinement by means of more accurate estimation methods.

Many different tools have been developed within the CRECK research activity. The OpenSMOKE++ [10] code is used to perform kinetic simulations of typical facilities such as jet stirred and flow reactors, 1D-2D laminar flames, shock tubes and rapid compression machines. The variables of interest are typically ignition delay times or laminar flame speeds of fuel/oxidizer mixtures, fuel consumption, intermediate and product species formation/disappearance at specific conditions of temperature (T) and pressure (p). Experimental measurements, typically stored in ASCII, CSV, XML formats on remote servers are compared to outputs from numerical simulations.

Beyond classical graphical comparisons (i.e., those typically reported in publications) the “Curve Matching approach” [6] allows for an objective, quantitative and automatic evaluation of model capability of predicting the variables of interest. If the model provides satisfactory agreement, subsequent steps of

[4] http://creckmodeling.chem.polimi.it/

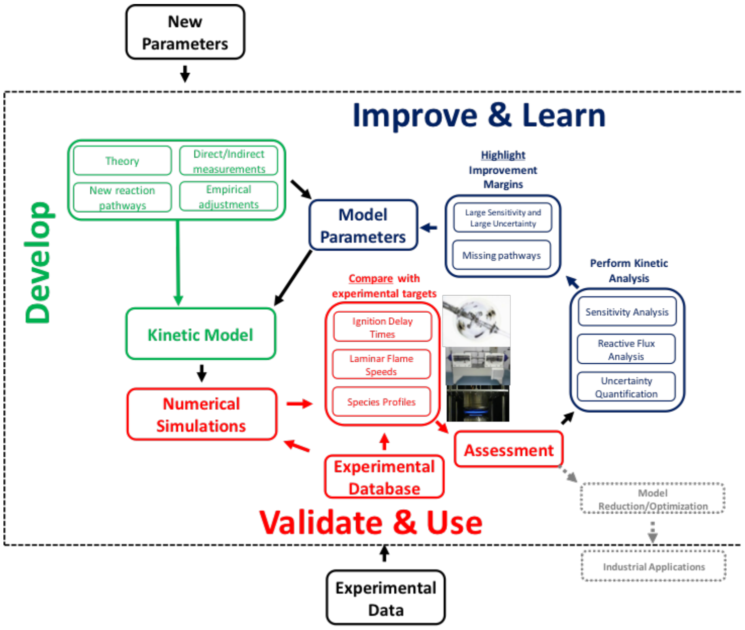

Fig. 1. Standard development, validation and refinement procedure of a chemical kinetic model for combustion applications.

optimization and reduction [26] make the model suitable for large scale computations of interest for industry. On the contrary, if the model shows deviations outside of the experimental uncertainties, relevant pathways can be identified by means of analysis tools and model parameters are further refined with better estimates. Indeed, the recent developments coupling high performance computing and theoretical chemistry allow the automatic generation of highly accurate parameters [8].

While the efficient integration of the above tools in a fully automatized system is one of today's challenges in kinetic modelling [18], efficient and smart data collection, formatting, analysis, conversion and storage is the new frontier for the domain.

The exponential growth in the number and complexity of scientific information in the combustion community (experimental data, models, theoretical investigations etc.) and the improved accuracy of experimental techniques and theoretical methods can be beneficial at best only if coupled with extremely efficient tools for acquiring, storing and analyzing of such information, thereof allowing real advances in knowledge.

Several initiatives to enable effective and structured data collection of experimental data for combustion science are available in the literature at present. Starting from the pioneering work of M. Frenklach and co-workers developing the PrIMe [15, 3] database which is still under continuous update at DLR

Stuttgart, the ReSpecTh repository largely improved and extended the previous approach by means of a more flexible, detailed and user-friendly structure [4, 27]. At present ReSpecTh collects into XML files $\sim$ 104 datasets ($\sim$ 105 data points) of relevance for the validation of combustion models. Despite the total number of combustion experiments is difficult to estimate, the extent of this collection is expected to increase more than linearly in the years to come. Additional data repositories such as ChemKED (Oregon State/Connecticut) [1, 29] and CloudFlame (KAUST) [2] further extended the interest in better structuring experimental information and increased the number of experimental data systematically collected through machine readable formats. Interestingly, the COST Action CM 1404 (SMARTCATs)[5] established a task force of scientist aiming at defining standards for data collection, allowing easy and effective coupling with the above systems.

On top of the reference repositories mentioned above, one should consider a similar amount of experimental information stored in a less structured format into many institutional servers belonging to experimental or modelling groups working in the field of combustion. As an example, the CRECK repository, which is taken as the basis for defining the requirements illustrated in this paper, is the result of data collection in $\sim$ 30 years of research efforts in modelling combustion phenomena. While the previous less systematic approach to data management relied on manual extraction and classification into spreadsheets or ASCII/txt files, a more accurate and very recent implementation relies on a relational database (MySQL) [21] with a structured interface for extracting experiments, and a collection of related files. In its "beta" version it contains references to 30 scientific papers and their associated experimental data ($\sim$ 60 datasets and $\sim$ 1000 data points). This tool is interoperable with the ReSpecTh repository and coupled with the OpenSMOKE++ suite of programs [10] for numerical replication of experimental data aiming at combustion models validation.

3 Dynamic analysis requirements

Starting from the scenario described in the previous Section and wishes described in Section 1, a set of requirements has been formulated.

The goal of these requirements is to enhance the efficiency and the effectiveness of data management for this context.

Different analysis can be performed on the collected data, independently from one another. The analysis requirements are presented and discussed in the following paragraphs.

3.1 Continuous multi-source integration

The need for a continuous multi-source integration comes from the variability of the information sources, which could vary over time and cannot, therefore, be presumed a priori. This integration has many facets, most notably:

[5] http://www.smartcats.eu/

- The *format* of the data, which could vary.
- The *semantic*, since the same concepts could be described differently in different sources and datasets. This requires an ontology-based semantic layer — which is dynamic itself — and a conceptualization of the information already stored.
- The *information* conveyed by the data, which can be related to different types of experiments and settings and can largely vary in terms of *accuracy*, *precision* and *coverage*, from experimental uncertainty, parameters needed for correct simulations and/or replicability of the same measurement.

A continuous management of the already-acquired information is necessary in order to update data already stored according to new requirements for the analysis. The need for updates is also related to the *information quality* (IQ) management: as new data and metadata are acquired or generated through processing, the IQ — with each single dimension that defines it — evolves. Therefore, for example, the data associated to an experiment may have a certain accuracy (which impacts on the overall quality), but further acquired information and/or processing can improve it without changes in the data itself. Moreover, complex data not only are characterized by objects which change over time, enriching their information, but also by explicit (*complex networks*) or implicit (*articulated objects*) relationships of interdependencies among objects. The IQ of such articulated objects is a function of the information quality of the sub-objects and of the other objects for which a relationship exists. Indeed, besides managing the quality of raw data, which is a problem addressed in the literature, the focus is on introducing the management of the quality of complex information through their relationship. Such cases are typically characterized by *context-dependent* information and these dependencies are in general not simply additive. For example, there could be "partial views" expressed by sub-entities which bring to a meaningful information only when combined.

An open issue is represented by the lack of a complete domain ontology. To face this challenge, a solution could be the automatic generation of ontological relationships based on the acquired data and *data mining* techniques. For example, finding synonyms for the same entities starting from papers text [16] or attributes related to an experiment through clustering and other machine learning techniques on the data available.

3.2 Dynamic acquisition

The requirement for the system is to continuously "find" and integrate new data automatically. This can be accomplished by a dynamic acquisition driven by the already stored data. The goal is to (potentially) enhance the IQ of the already stored data by finding new information about them or new related data (for example, new experiments for a stored model).

Given the continuous validation performed on the stored information, the acquisition ultimately aims at better assessing the IQ and in general enhancing the data coverage.

This is accomplished by extending the concept of “focused crawling” [31]. The best predicate for querying, in general, changes over time and depends on a background knowledge. Acquired data can drive the acquisition of new data in a *virtuous cycle*. The background knowledge is certainly composed by the already acquired information, but also by the list of preceding queries and their results. Indeed, when acquiring data from unreliable sources it is not possible to make strong assumptions about them and an *exploratory approach* must be employed (see the discussion in Subsection 3.4).

Since the goal is to acquire as much information as possible, different source types must be taken into account. In particular, there are *structured* and *unstructured* source types. While the first include repositories or manual inputs and are handled mainly through the integration process illustrated in Subsection 3.1, the second includes valuable sources like papers and web pages without or with little structured information. This requires to *extract* information from unstructured text and images, using mining techniques (e.g. [11, 23]) and image analysis (e.g. [20]). Semi-automatic techniques can also be employed (e.g. [17]).

3.3 Continuous validation

The process of continuous validation entails the matching of already stored information with new information as it is acquired.

This requirement mainly comes from the need of validating *models* and *experimental data*, one respect to the others through curve matching techniques, as described in Section 2, but could be extended also to validations based on other kinds of data and metadata, e.g., authors, experiment types, and information quality.

Validation is performed through *cross-comparisons* which require efficient means of *extracting* the right data, *comparing* them taking into account the differences and the lacks that could exist in their representations and *enriching* the entities (models, experimental data, etc.) with the results.

This is a continuous process: for example, a model must be tested against new experimental data as they become available over time and therefore the validity of the model itself evolves over time.

Validating means also *verifying* data and models. However, there are many different situations that must be handled. For example:

- There could be a set of models that fails in a particular condition because there are no data for that condition and therefore new data should drive the refinement of models (“experimental design”).
- There could be data for an experiment that is not congruent with other data for the same experiment and therefore should be repeated and verified.
- There could be that *all* models fail for an experiment because they are ignoring something. In this case the experiment is correct and instead the models should be improved.

3.4 Data exploration

Articulated (meta)data can provide support for interactive and evolving data exploration [13].

This has to do with the need of automatically or manually querying for information which are in general incomplete, heterogeneous or may not exist at all.

In particular, manual interventions are key to maintain an overall high IQ resolving conflicts and enriching domain knowledge, and they need to be supported by effective query techniques. These include summarization [9], result refinement, iterative exploration [28] and a *top-k* query processing environment [25], thus improving the returned "manageable" results.

4 Proposed architecture

After listing the analysis requirements in Section 3, this section outlines an initial architecture to support them, focusing on architectural requirements.

The main components are sketched in Figure 2.

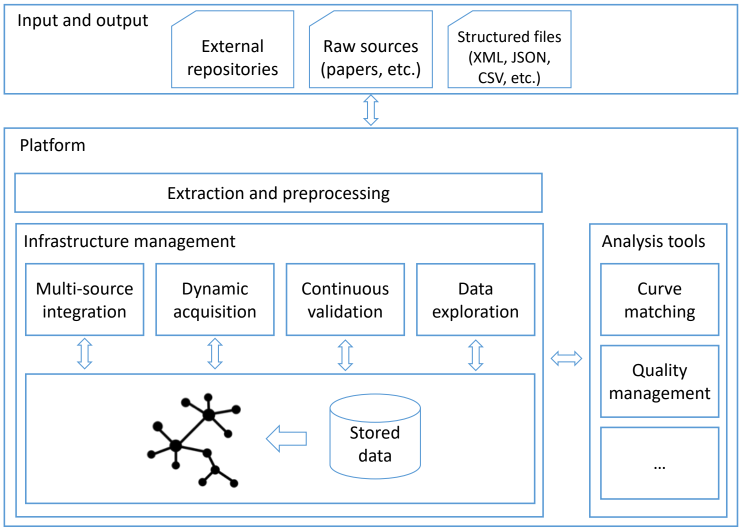

Fig. 2. Sketch of an initial architecture for the infrastructure.

A service-based approach has two advantages: the integration of legacy components with new components and the autonomous development of each of the services. Analysis tools need to be wrapped in services uniforming their interfaces to interact with infrastructure management components.

All the services access a common database. The database should optimize the functionalities of the components.

This can be obtained first of all selecting *knowledge graph* as conceptual-modeling. RDF allows to model relationships among objects, necessary for modeling interdependencies among entities (required by the continuous multi-source integration) and for data exploration — also through faceted search [22] — which is key in all the requirements. Semantic graphs can also be exploited for mining activities [22] and to model varying precision and accuracy. Graph modeling can be accomplished through a graph database [19] or via mappings that enable graph reasonings over traditional relational databases [7, 24].

Another important requisite for the database is *indexing* to efficiently search and retrieve data based on the patterns of the acquisition/validation/exploration cycle and the metadata. For example, to validate an experiment, all the models for that particular experiment should be retrieved. This is also related to *top-k* query processing for data exploration.

Latency should be limited by parallelizing and putting in background tasks which not require to be responsive.

Finally, given the variety of existing formats and standards, the import/export interfaces should provide conversion tools.

5 Concluding remarks

The effective integration of data science in current approaches and techniques for science and engineering is one of todays societal challenges, potentially allowing rapid and extremely significant technological advancements. Despite the generality of this perspective, this work focused on the well defined domain of combustion science, typically dealing with the investigation and development of new, cleaner and more efficient combustion technologies (i.e. fuel and engines). Even though a consistent delay exists compared to other industrial or research areas (e.g., pharmaceutics, organic chemistry etc.), recent implementation of data repositories specifically conceived for combustion kinetic modelling activities [29, 3, 2, 4], ongoing initiatives within the community and incentives from the EU, further encourage the activity outlined in this work. After a qualitative analysis of the domain, requirements for the dynamic analysis of the large amount of information available have been discussed, focusing on the continuous multi-source integration, the dynamic acquisition, the continuous validation and data exploration, with the related issues. Finally, a preliminary architecture has been defined, setting the basis for its implementation, extension and refinements in future activities.

References

1. ChemKED repository. http://www.chemked.com/.
2. CloudFlame repository. https://cloudflame.kaust.edu.sa/.
3. PrIMe repository. http://primekinetics.org/.
4. ReSpecTh repository. http://respecth.hu/.

5. J. M. Bergthorson and M. J. Thomson. A review of the combustion and emissions properties of advanced transportation biofuels and their impact on existing and future engines. *Renewable and sustainable energy reviews*, 42:1393–1417, 2015.
6. M. Bernardi, M. Pelucchi, A. Stagni, L. Sangalli, A. Cuoci, A. Frassoldati, P. Secchi, and T. Faravelli. Curve matching, a generalized framework for models/experiments comparison: An application to n-heptane combustion kinetic mechanisms. *Combustion and Flame*, 168:186–203, 2016.
7. D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, and G. Xiao. Ontop: Answering sparql queries over relational databases. *Semantic Web*, 8(3):471–487, 2017.
8. C. Cavallotti, M. Pelucchi, S. Klippenstein, and EStokTP. Electronic structure to temperature and pressure dependent rate constants. *unpublished*, 2017.
9. A. Cohan and N. Goharian. Scientific article summarization using citation-context and article's discourse structure. *arXiv preprint arXiv:1704.06619*, 2017.
10. A. Cuoci, A. Frassoldati, T. Faravelli, and E. Ranzi. Opensmoke++: An object-oriented framework for the numerical modeling of reactive systems with detailed kinetic mechanisms. *Computer Physics Communications*, 192:237–264, 2015.
11. V. Daudaravicius. A framework for keyphrase extraction from scientific journals. In *International Workshop on Semantic, Analytics, Visualization*, pages 51–66. Springer, 2016.
12. A. de Waard. Research data management at elsevier: Supporting networks of data and workflows. *Information Services & Use*, 36(1-2):49–55, 2016.
13. N. Di Blas, M. Mazuran, P. Paolini, E. Quintarelli, and L. Tanca. Exploratory computing: a comprehensive approach to data sensemaking. *International Journal of Data Science and Analytics*, 3(1):61–77, 2017.
14. C. Francalanci, B. Pernici, and G. Scalia. Exploratory spatio-temporal queries in evolving information. In C. Doulkeridis, G. A. Vouros, Q. Qu, and S. Wang, editors, *Mobility Analytics for Spatio-Temporal and Social Data*, pages 138–156, Cham, 2018. Springer International Publishing.
15. M. Frenklach. Transforming data into knowledgeprocess informatics for combustion chemistry. *Proceedings of the combustion Institute*, 31(1):125–140, 2007.
16. K. Gábor, H. Zargayouna, I. Tellier, D. Buscaldi, and T. Charnois. A typology of semantic relations dedicated to scientific literature analysis. In *International Workshop on Semantic, Analytics, Visualization*, pages 26–32. Springer, 2016.
17. D. Jung, W. Kim, H. Song, J.-i. Hwang, B. Lee, B. Kim, and J. Seo. Chartsense: Interactive data extraction from chart images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6706–6717. ACM, 2017.
18. M. Keceli, Y.-P. L. Elliott, M. Johnson, C. Cavallotti, Y. Georgievskii, W. P. Green, J. Wozniak, A. M. Jasper, and S. Klippenstein. Automated computational thermochemistry for butane oxidation: A prelude to predictive automated combustion kinetic. 2018.
19. L. Libkin, W. Martens, and D. Vrgoč. Querying graphs with data. *Journal of the ACM (JACM)*, 63(2):14, 2016.
20. J. Poco and J. Heer. Reverse-engineering visualizations: Recovering visual encodings from chart images. In *Computer Graphics Forum*, volume 36, pages 353–363. Wiley Online Library, 2017.
21. A. Rigamonti. Automatic modeling system: a database based infrastructure to develop, validate and evaluate scientific models. an application to combustion kinetic models, 2017.

22. P. Ristoski and H. Paulheim. Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, 36:1–22, 2016.
23. F. Ronzano and H. Saggion. Knowledge extraction and modeling from scientific publications. In *International Workshop on Semantic, Analytics, Visualization*, pages 11–25. Springer, 2016.
24. A. Schätzle, M. Przyjaciel-Zablocki, S. Skilevic, and G. Lausen. S2rdf: Rdf querying with sparql on spark. *Proceedings of the VLDB Endowment*, 9(10):804–815, 2016.
25. M. A. Soliman, I. F. Ilyas, and K. C.-C. Chang. Top-k query processing in uncertain databases. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 896–905. IEEE, 2007.
26. A. Stagni, A. Frassoldati, A. Cuoci, T. Faravelli, and E. Ranzi. Skeletal mechanism reduction through species-targeted sensitivity analysis. *Combustion and Flame*, 163:382–393, 2016.
27. T. Varga, T. Turányi, E. Czinki, T. Furtenbacher, and A. Császár. Respecth: a joint reaction kinetics, spectroscopy, and thermochemistry information system. In *Proceedings of the 7th European Combustion Meeting*, volume 30, pages 1–5, 2015.
28. A. Wasay, M. Athanassoulis, and S. Idreos. Queriosity: Automated data exploration. In B. Carminati and L. Khan, editors, *2015 IEEE International Congress on Big Data, New York City, NY, USA, June 27 - July 2, 2015*, pages 716–719. IEEE, 2015.
29. B. W. Weber and K. E. Niemeyer. Chemked: A human-and machine-readable data standard for chemical kinetics experiments. *International Journal of Chemical Kinetics*, 2017.
30. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.
31. R. Yu, U. Gadiraju, B. Fetahu, and S. Dietze. Adaptive focused crawling of linked data. In *International Conference on Web Information Systems Engineering*, pages 554–569. Springer, 2015.