


# Conference Linked Data: The ScholarlyData Project

Andrea Giovanni Nuzzolese<sup>1</sup>, Anna Lisa Gentile<sup>2</sup>, Valentina Presutti<sup>1</sup>,  
and Aldo Gangemi<sup>1,3</sup>

<sup>1</sup> Semantic Technology Lab, ISTC-CNR, Padova, Italy  
`andrea.nuzzolese@istc.cnr.it`, `{valentina.presutti,aldo.gangemi}@cnr.it`

<sup>2</sup> Data and Web Science Group, University of Mannheim, Mannheim, Germany  
`annalisa@informatik.uni-mannheim.de`

<sup>3</sup> LIPN, Université Paris 13, Sorbone Cité, UMR CNRS, Paris, France

**Abstract.** The Semantic Web Dog Food (SWDF) is the reference linked dataset of the Semantic Web community about papers, people, organisations, and events related to its academic conferences. In this paper we analyse the existing problems of generating, representing and maintaining Linked Data for the SWDF. With this work (i) we provide a refactored and cleaned SWDF dataset; (ii) we use a novel data model which improves the Semantic Web Conference Ontology, adopting best ontology design practices and (iii) we provide an open source workflow to support a healthy growth of the dataset beyond the Semantic Web conferences.

**Permanent URL:** <https://w3id.org/scholarlydata>

**Resource type:** Ontology and dataset.

## 1 Introduction

A good practise in the Semantic Web community is to encourage the publication of Linked Data about scientific conferences in the field, as a way of “eating our own dog food” [8]. The main example is the *Semantic Web Dog Food*<sup>1</sup> (SWDF), a corpus that collects Linked Data about papers, people, organisations, and events related to academic conferences. Currently, all main Semantic Web conferences and related events publish their data as Linked Data on SWDF, but for many other conferences, events and publication venues information is still not available in a structured and linked form. On the other hand the growth of available content with respect to the early times of SWDF poses data management issues and reveals design problems which were not foreseen when the dataset was at its initial stage. There are several challenges to pursue the maintenance of a healthy and sustainable SWDF for the future: (i) the availability of appropriate vocabularies to express the current state of the data; (ii) the shared knowledge

---

<sup>1</sup> SWDF: <http://data.semanticweb.org>.

of such vocabularies; (iii) the availability of tools to ease the task of data acquisition, conversion, integration, augmentation, verification and finally publication; (iv) the ongoing maintenance of the dataset.

In this work we address these issues and we propose a refactoring of the Semantic Web Conference (SWC) Ontology<sup>2</sup>. The new ontology, named *conference-ontology* [12], adopts best ontology design practices (e.g. Ontology Design Patterns, ontology reuse and interlinking) and guarantees interoperability with SWC ontology and all other pertinent vocabularies. We use cLODg<sup>3</sup> (conference Linked Open Data generator) [6] to regenerate the SWDF dataset according to *conference-ontology* and provide a sustainable solution for the growth of the dataset in the future.

The main advantage of the proposed approach is the availability of a shared procedure and open source tools for conference data generation, with the primary goal to ensure the sustainability and usability of our own Semantic Web Dog Food and the ease of data contribution from beyond our community. We make the new resource available at <https://w3id.org/scholarlydata> as data dump, SPARQL endpoint and we offer the facilities to generate data about new conferences using cLODg and submit it for addition to *scholarlydata*. The newly submitted data is manually checked before inclusion to avoid corruption of the dataset and general spam.

## 2 State of the Art

The first considerable effort to offer comprehensive semantic descriptions of conference events is represented by the metadata projects at ESWC 2006 and ISWC 2006 conferences [11], with the Semantic Web Conference Ontology<sup>4</sup> being the vocabulary of choice to represent such data. Increasing number of initiatives are pursuing the publication about conferences data as Linked Data, mainly promoted by publishers such as Springer<sup>5</sup> or Elsevier<sup>6</sup> amongst many others. For example, the knowledge management of scholarly products is an emerging research area in the Semantic Web field known as Semantic Publishing [14]. Semantic Publishing aims at providing access to semantic enhanced scholarly products with the aim of enabling a variety of semantically oriented tasks, such as knowledge discovery, knowledge exploration and data integration. The Semantic Publishing challenge [9] is a breakthrough in this direction. Its objective is assessing the quality of systems that extract meaningful metadata from scholarly articles and represent them as RDF. Similarly, the *Jailbreaking the PDF* initiative [5] is aimed at creating a formal flexible infrastructure to extract semantic

<sup>2</sup> [http://data.semanticweb.org/ns/swc/swc\\_2009-05-09.html](http://data.semanticweb.org/ns/swc/swc_2009-05-09.html).

<sup>3</sup> cLODg is an Open Source tool that provides a formalised process for the conference metadata publication workflow <https://github.com/anuzzolese/cLODg2>.

<sup>4</sup> Semantic Web Conference Ontology [http://data.semanticweb.org/ns/swc/swc\\_2009-05-09.html](http://data.semanticweb.org/ns/swc/swc_2009-05-09.html).

<sup>5</sup> <http://lod.springer.com/wiki/bin/view/Linked+Open+Data/About>.

<sup>6</sup> <http://data.elsevier.com/documentation/index.html>.

information from PDF documents as domain-specific annotations. Despite these continuous efforts, it has been argued that lots of information about academic conferences is still missing or spread across several sources in a largely chaotic and non-structured way [1]. Besides the problem of missing content, one of the other major challenges with scholarly data is to ensure data quality, which means dealing with data-entry errors, disparate citation formats, lack of (enforcement of) standards, imperfect citation-gathering software, ambiguous author names and abbreviations of publication venue titles [10]. Currently the generation of data for the SWDF corpus still relies on little or no strategies to deal with duplicates, inconsistencies, misspelling and name variations. In this work we aim to close these gaps by making available a solid data model and a shared and open workflow (available Open Source) as a long term solution for the population and maintenance of an enhanced version of the SDWF dataset.

### 3 The SWDF and Its Current Issues

The SWDF uses the Semantic Web Conference (SWC) ontology as the reference ontology for modelling data about academic conferences. The SWC ontology combines existing widely accepted vocabularies (i.e. FOAF<sup>7</sup>, SIOC<sup>8</sup> and Dublin Core<sup>9</sup>) and relies on the SWRC<sup>10</sup> (Semantic Web for Research Communities) ontology for modelling entities such as accepted papers, authors, their affiliations, talks and other events, the organising committee and all other roles involved. The core types of SWC ontology are `foaf:Person` for describing people, `foaf:Organization` for organisations (e.g. universities, research institutions, etc.), `swc:Artefact` for documents (e.g. papers, proceedings, etc.), `swc:OrganisedEvent` for events and `swc:Role` for the people roles at the conference. Unfortunately, the lack of clear guidelines for data generation and maintenance and some modelling choices of the SWC ontology affect the current quality of SWDF. The data generation is based on a collaborative model that delegates the metadata chairs of each conference to independently deal with the process of generating conference Linked Data. Linked Data are generated from a variety of formats typically provided by a conference management system (e.g. EasyChair). While the collaborative process is beneficial to the growth of the dataset and its adoption in the community, the lack of clear guidelines and of standard tools supporting the generation process affects the quality of generated data. Examples are: (i) a portion of the included conference/workshop data use vocabularies or ontologies which are not aligned to the SWC ontology and, in some cases, no longer maintained or existing (e.g. `swrc-ext`<sup>11</sup> or `xmlondon`<sup>12</sup>); (ii) the usage of classes and properties not defined

<sup>7</sup> <http://xmlns.com/foaf/spec/>.

<sup>8</sup> <http://rdfs.org/sioc/spec/>.

<sup>9</sup> <http://dublincore.org/documents/dcmi-terms/>.

<sup>10</sup> [http://ontoware.org/swrc/swrc/SWRCOWL/swrc\\_updated\\_v0.7.1.owl](http://ontoware.org/swrc/swrc/SWRCOWL/swrc_updated_v0.7.1.owl).

<sup>11</sup> [http://www.cs.vu.nl/~mcaklein/onto/swrc\\_ext/2005/0](http://www.cs.vu.nl/~mcaklein/onto/swrc_ext/2005/0).

<sup>12</sup> <http://xmlondon.com/ns/swc/ontology>.

in the SWC ontology and introduced without providing an extension of the ontology (e.g. `swc:room`, `swc:editorList`, `swc:completeGraph`, `swc:IW3C2Liaison`, `swc:SemanticWebTechnologiesCo-ordinator`, etc.); (iii) the misuse of properties (either defined in the SWC ontology or in other vocabularies/ontologies) with respect to their domain and range; (iv) typos (e.g. the materialisation of triples having the predicate `swc:partOf` instead of `swc:isPartOf`). In addition, we argue that the SWC ontology itself has intentional issues, mainly concerning the modelling of *affiliations*, *roles* and *lists*. *Affiliations* (of people to organisation) are represented via the object property `swrc:affiliation` from the SWRC ontology while the membership relation (organisation to people) via the property `foaf:member`. Although intuitive, this representation ignores the temporal dimension (i.e. the time when a given affiliation is held by an actor) that is relevant to interpret affiliations correctly. For example, with this model it is not possible to provide a correct answer to a simple competency question, such as “What was the affiliation of a person when participating to a certain conference?”. *Roles* such as program chair, track chair, etc. are currently modelled using an ontology pattern based on the reification of a  $n$ -ary relation. The  $n$ -ary relation is identified by individuals of the class `swc:Role` which are used to associate people to events. The SWC ontology contains a very basic set of role classes (i.e. `swc:Chair`, `swc:Delegate`, `swc:Presenter` and `swc:ProgrammeCommitteeMember`) represented as sub-classes of `swc:Role`. This choice allows to instantiate the small set of different Role classes and cover the roles at specific events. For example, instead of sub-classing the `swc:Chair` class with `MainChair`, `WorkshopChair`, `TutorialChair`, etc., the different types of chairs should simply be instances of the generic `swc:Chair` and labelled appropriately (e.g. `iswc2015:general-chair`<sup>13</sup>). The problem with this solution is that the (individuals representing) roles are defined locally to each conference, e.g. there is a different individual for representing the role “general chair” for each conference in the dataset. This causes the presence of 1,717 distinct individuals in the current dataset that truly represent a set of only 34 unique roles (cf. Sect. 4). Hence, it is difficult to answer simple queries like “Who was the general chair at each edition of ISWC?” without using regular expressions on roles’ labels (such labels are heterogeneous and not always provided). Finally, *lists of authors* are represented via the property `bibo:authorList`, which accepts `rdf:List` or `rdf:Seq` as range. Therefore lists of authors in the SWDF are expressed via the properties `rdf:_1`, `rdf:_2`, `rdf:_3`, etc., based on `rdfs:ContainerMembershipProperty`. This solution makes querying and reasoning on ordered list of authors very hard [2].

## 4 A Sustainable SWDF

Our solution for enhancing the SWDF and solving the issues described in Sect. 3 is based on (i) the refactoring of the SWC ontology, (ii) the refactoring of the

<sup>13</sup> The prefix `iswc2015:` stands for the namespace <http://data.semanticweb.org/conference/iswc/2015/>.

current SWDF dataset and (iii) a fully implemented open source workflow to generate, verify and add data to SWDF. The proposed refactoring of the SWC ontology, *conference-ontology*<sup>14</sup>, is a new self-contained ontology, which exploits Ontology Design Patterns (ODP) [3]. We model *affiliations* reusing the time indexed situation ODP<sup>15</sup> and the *roles* held by people at a conference with the time indexed person role ODP<sup>16</sup>. Both patterns provide commonly accepted solutions to model complex situations as n-ary relations, amongst many other available ones [4].

The classes `conf:AffiliationDuringEvent` and `conf:AffiliationAtTimeOfSubmission` model situations where a person (an individual of the class `conf:Person`) is affiliated to an organisation (an individual of the class `conf:Organisation`) at a specific time, which can be either an interval (coinciding with the conference dates) or the instant when the paper was submitted. This allows the representation of cases where a person changes affiliation in the time interval between paper submission and conference event. Similarly, the class `conf:RoleDuringEvent` associates a person with a role (an individual of the class `conf:Role`) at a conference. Additionally, `conf:AffiliationDuringEvent` and `conf:AffiliationAtTimeOfSubmission` can be associated with `conf:AffiliationRole`, a subclass of `conf:Role`, to represent the role held by a person within an organisation. We reused the Sequence ODP<sup>17</sup> to represent ordered *lists of authors*. We represent a list with `conf:List`, whose items are individuals of the class `conf:ListItem`. The association between `conf:ListItem` and `conf:List` is done via the property `conf:isItemOf`. A `conf:List` has pointers to the first (`conf:hasFirstItem`) and the last item (`conf:hasLastItem`). Each `conf:ListItem` is linked to its predecessor (`conf:hasPreviousItem`) and successor (`conf:hasNextItem`). This new modelling overcomes the limitation in the current SWDF offering a new range of services for scholarly monitoring, such as statistics on career development, change of affiliations over time, covered roles at conferences in order to monitor their involvement and impact at different granularity levels, ranging from a broader scientific area to specific communities or conferences. An example of query to obtain all roles covered overtime by a specific researcher is the following:

```
PREFIX person: <http://w3id.org/scholarlydata/person/>
PREFIX conf: <http://w3id.org/scholarlydata/ontology/conference-ontology.owl#>
SELECT ?role ?during
WHERE{
  person:valentina-presutti conf:holdsRole ?roleAt .
  ?roleAt conf:withRole ?role .
  ?roleAt conf:during ?during}
```

To guarantee interoperability with SWC and all other already used vocabularies in the SWDF dataset, we produced extensive alignments<sup>18</sup>, which allow the materialisation of triples via reasoning. We include alignments to:

<sup>14</sup> The ontology diagram, description and specification are available at <http://w3id.org/scholarlydata/ontology>.

<sup>15</sup> <http://ontologydesignpatterns.org/cp/owl/timeindexedsituation.owl>.

<sup>16</sup> <http://ontologydesignpatterns.org/cp/owl/timeindexedpersonrole.owl>.

<sup>17</sup> <http://ontologydesignpatterns.org/cp/owl/sequence.owl>.

<sup>18</sup> <http://w3id.org/scholarlydata/ontology/conference-ontology-alignments.owl>.

- the SWC ontology itself to guarantee backward interoperability with SWDF;
- the top level classes of Dolce D0<sup>19</sup> for interoperability with a series of linked datasets aligned to it (e.g. DBpedia);
- all relevant SPAR ontologies<sup>20</sup> such as: FaBIO [13] for compliance with FRBR; DoCO for modelling the part relations (`conf:hasPart` and its inverse `conf:isPartOf` existing between abstracts `conf:Abstract`, articles `conf:-InProceedings` and the books of proceedings `conf:Proceedings`); PRO and SCORO for modelling roles as defined in SPAR;
- the Organization Ontology<sup>21</sup> for modelling organisations, roles and affiliations;
- FOAF for modelling people;
- SKOS<sup>22</sup> for modelling broader/narrower relations;
- ICATZD<sup>23</sup> for events;
- the Collection Ontology [2] for modelling the sequences represented by the lists of authors.

## 5 Scholarlydata.org

Using cLODg and our new *conference-ontology* we performed a batch cleaning of the whole SWDF dataset, consisting of 48 conferences and 235 workshops. The new dataset contains 93,519 individuals. The distribution of classes is reported in Table 1.

**Table 1.** Number of unique individuals for each class of *conference-ontology* generated with cLODg.

Type	Individuals	Type	Individuals
<code>conf:TimeIndexedSituation</code>	20,998	<code>conf:RoleDuringEvent</code>	6,510
<code>conf:ListItem</code>	14,805	<code>conf:List</code>	4,463
<code>conf:AffiliationDuringEvent</code>	14,488	<code>conf:InProceedings</code>	4,393
<code>conf:Agent</code>	12,490	<code>conf:OrganisedEvent</code>	2,882
<code>conf:Person</code>	9,682	<code>conf:Organisation</code>	2,808

For the role definitions we corrected the current 1,717 roles in SWDF, defined at conference level, by generating 34 roles at global level and reusing them at conference level. E.g. the role `role:general-chair`<sup>24</sup> is one individual which can be reused in all conferences with the relation `conf:withRole`. These 34 roles

<sup>19</sup> <http://www.ontologydesignpatterns.org/ont/dul/d0.owl>.

<sup>20</sup> <http://www.sparontologies.net/ontologies>.

<sup>21</sup> <https://www.w3.org/TR/vocab-org/>.

<sup>22</sup> <https://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/>.

<sup>23</sup> <http://www.w3.org/2002/12/cal/icaltzd>.

<sup>24</sup> The prefix *role:* stands for the namespace <http://w3id.org/scholarlydata/role/>.

are organised in a hierarchy by using SKOS to express broader and narrower relations between them, e.g. the role `role:chair` is defined as `skos:broader role:general-chair`. The current list of roles can be obtained using the query:

```
PREFIX conf: <http://w3id.org/scholarlydata/ontology/conference-ontology.owl#>
SELECT distinct ?role
WHERE{ ?person conf:holdsRole ?roleAt .
?roleAt conf:withRole ?role }
```

Using cLODg to produce metadata about a new conference guarantees that pertinent roles are reused if already existing the dataset.

We produced instance level alignments of (i) individuals of `conf:Person` to ORCID<sup>25</sup> (Open Researcher and Contributor ID) and (ii) individuals of `conf:InProceedings` to DOI<sup>26</sup> (Digital Object Identifier), whenever possible. ORCID provides persistent digital identifiers for scientific researchers and academic authors. A DOI is a serial code used to uniquely identify digital objects, particularly used for electronic documents. The alignments to ORCID were produced by using the public API provided by ORCID<sup>27</sup>. The references to DOI were produced by using the API provided by Crossref<sup>28</sup>, performing a search on each article title.

All data is uploaded on <https://w3id.org/scholarlydata> where can be accessed in different formats (i.e. HTML+RDFa, RDF/XML, Turtle, N-TRIPLES, and JSON-LD) via URI dereferencing, queried via SPARQL or downloaded as single RDF dumps for each conference and workshop. Each dump is provided in two versions: a simple one, where data is represented by the *conference-ontology* only and one containing all the alignments (and therefore also complaint to SWDF), which have been materialised using a reasoner. These dumps are released with the “creative commons by 3.0” license<sup>29</sup> and are described by using the VOID vocabulary<sup>30</sup>. Additionally, we explicitly state the primary source of our data is the SWDF by using the property `prov:hadPrimarySource` of PROV-O<sup>31</sup>. Dump data is also publicly available on datahub<sup>32</sup>. It is worth remarking that cLODg is released as an open source software with the MIT License<sup>33</sup> and can be used by metadata curator to add data about a new conference. In fact, cLODg provides a nearly one-click process to produce conference Linked Data and includes all the components for data transformation, deduplication, URIs reuse, alignment of individuals to external resources, etc. and assures that data is produced according to the *conference-ontology* and compliant with the SWDF. An early description of cLODg can be found in [7] and in the github repository for its

<sup>25</sup> <http://orcid.org/>.

<sup>26</sup> <http://www.doi.org/index.html>.

<sup>27</sup> <http://members.orcid.org/api/introduction-orcid-public-api>.

<sup>28</sup> <http://www.crossref.org/guestquery/>.

<sup>29</sup> <http://creativecommons.org/licenses/by/3.0>.

<sup>30</sup> <http://vocab.deri.ie/void>.

<sup>31</sup> <https://www.w3.org/TR/prov-o>.

<sup>32</sup> <https://datahub.io/dataset/scholarlydata>.

<sup>33</sup> <https://opensource.org/licenses/MIT>.

newer version<sup>34</sup>. By providing a user friendly data generation tool we aim at encouraging the growth of the dataset beyond the Semantic Web community.

## 6 Conclusions and Future Work

This paper analyses the Semantic Web Dog Food dataset and discusses its quality and sustainability issues. As the main scholarly dataset for the Semantic Web community, we believe it is important that the dataset is maintained in good health. We therefore perform a refactoring on the dataset addressing its current issues and we make the cLODg workflow publicly available as potential solution for future maintenance. The new resource <https://w3id.org/scholarlydata> is publicly available both as dump download and SPARQL endpoint, with facilities to upload new data. With the availability of cLODg as standard Linked Data publication workflow, we believe that *scholarlydata* has the potential to grow way beyond the Semantic Web conferences. As future work we plan a systematic evaluation of the resource and the introduction of more sophisticated components to deal with instance matching in the cLODg workflow. Moreover we will work on fostering collaboration with Conference Management System providers, to provide cLODg as a build-in facility in the systems.

## References

1. Bryl, V., Birukou, A., Eckert, K., Kessler, M.: What is in the proceedings? combining publishers and researchers perspectives. In: Proceedings of SePublica 2014, Anissaras, Greece, May 25th 2014
2. Ciccicarese, P., Peroni, S.: The collections ontology: creating and handling collections in OWL 2 DL frameworks. *Semant. Web* **5**(6), 515–529 (2014)
3. Gangemi, A., Presutti, V.: Ontology design patterns. In: Staab, S., Studer, R. (eds.) *Handbook on Ontologies*, 2nd edn. Springer, Heidelberg (2009)
4. Gangemi, A., Presutti, V.: A multi-dimensional comparison of ontology design patterns for representing  $n$ -ary relations. In: Emde Boas, P., Groen, F.C.A., Italiano, G.F., Nawrocki, J., Sack, H. (eds.) *SOFSEM 2013. LNCS*, vol. 7741, pp. 86–105. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-35843-2\\_8](https://doi.org/10.1007/978-3-642-35843-2_8)
5. Garcia, A., Murray-Rust, P., Burns, et al.: Pdfjailbreak-a communal architecture for making biomedical pdfs semantic. In: Proceedings of BioLINK SIG 2013, p. 13 (2013)
6. Gentile, A.L., Acosta, M., Costabello, L., Nuzzolese, A.G., Presutti, V., Recupero, D.R., Live, C.: Accessible and sociable conference semantic data. In: Proceedings of WWW 2015 (Companion Volume), pp. 1007–1012. ACM (2015)
7. Gentile, A.L., Nuzzolese, A.G.: cLODg - conference linked open data generator. In: Proceedings of the ISWC 2015 Posters and Demonstrations Track. CEUR-WS.org (2015)
8. Harrison, W.: Eating your own dog food. In: *Industrial, Organizational Psychology*, 5–7 June 2011

<sup>34</sup> <https://github.com/anuzzolese/cLODg2>.



9. Lange, C., Iorio, A.: Semantic publishing challenge – assessing the quality of scientific output. In: Presutti, V., et al. (eds.) *SemWebEval 2014*. CCIS, vol. 475, pp. 61–76. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-12024-9\\_8](https://doi.org/10.1007/978-3-319-12024-9_8)
10. Lee, D., Kang, J., Mitra, P., Giles, C.L., On, B.-W.: Are your citations clean? *Commun. ACM* **50**(12), 33–38 (2007)
11. Möller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for semantic web dog food: The eswc and iswc metadata projects. In: *Proceedings of ISWC'07/ASWC 2007*, pp. 802–815. Springer-Verlag, Berlin, Heidelberg (2007)
12. Nuzzolese, A.G., Gentile, A.L., Presutti, V., Gangemi, A.: Semantic web conference ontology - a refactoring solution. In: *The Semantic Web: ESWC 2016 Satellite Events*, Springer, Heidelberg (2016). page to appear
13. Peroni, S., Shotton, D.: FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *J. Web Semant.* **17**, 33–43 (2012)
14. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing* **22**(2), 85–94 (2009)