Exploring the Generation and Integration of Publishable Scientific Facts Using the Concept of Nano-publications

Amanda Clare[1,3], Samuel Croset[2,3] (croset@ebi.ac.uk), Christoph Grabmueller[2,3], Senay Kafkas[2,3], Maria Liakata[1,2,3], Anika Oellrich[2,3], and Dietrich Rebholz-Schuhmann[2,3]

[1] University of Aberystwyth
[2] European Bioinformatics Institute
[3] All authors contributed equally to this work

Abstract. Publication formats are being sought that facilitate automatic processing and knowledge integration and are better suited to the current pace of research. Here we present an infrastructure for producing and consuming minimal publishable units, nano-publications, directly from a researcher's electronic notes or manuscripts which allow the integration of multiple resources. We describe a feedback loop resulting from the use of nano-publications, give a detailed example, and explain how this can be combined with existing web technologies.

1 Introduction

With the ever growing amount of scientific literature the automatic analysis of scientific content has become crucial. As part of the research life-cycle, researchers constantly need to retrieve relevant documents, pull out facts, reuse and reference them. Yet, currently many scientific facts are 'buried' in the plethora of information contained in traditional scientific publications. Repositories, such as PubMed[4], store electronic versions of scientific publications but the scientific facts they contain are still not available for automated processing. Research in recent years has looked at the enhancement of scientific publications with semantic meta-data in order to facilitate information retrieval and information extraction from them. Numerous initiatives from the publishing world have been launched to address this task, such as Structured Digital Abstracts [3], The Royal Society of Chemistry (RSC) Project Prospect[5] and UKPMC[6], a major initiative focusing on the integration and alignment of literature with current knowledge resources and databases. An example system for the enhancement of documents with content from a variety of external resources is Utopia Documents [1]. Meanwhile, scientists have begun to seek more rapid and interactive ways of airing their findings and retrieving the findings of others, through media such

[4] http://www.ncbi.nlm.nih.gov/pubmed/
[5] http://www.rsc.org/Publishing/Journals/ProjectProspect/
[6] http://ukpmc.ac.uk/

as blogs and wikis. Wikis are collaborative tools, enabling users to collect, share and edit information while blogs are incremental content management systems enabling rapid publication of information. In traditional wiki systems and blogs, accessing, querying, retrieving and aggregating data is difficult since the knowledge is represented in unstructured form. These issues induced the emergence of semantic wiki systems like Semantic MediaWiki and DBpedia and commercial or semi-commercial web content providers (e.g. Apture). However, such tools are currently not aimed at scientists and do not offer the precision and level of detail that scientists need to make their work unambiguous and available in a machine-readable, reusable form to others. Importantly, it is not easy to receive credit for statements on wikis or blogs, or to cite the information therein as one would do with a standard publication. The nano-publication (NP) [6] has been proposed as a new form of academic publishing. Unlike other initiatives for linking shared statements in the literature [7], a NP is defined as a citable unit containing a set of annotated statements which capture knowledge in the form of Resource Description Framework (RDF) triples, representing three concepts (subject, predicate, object) [4]. The RDF graph emerging from the triples can be identified with a name, a procedure which was coined as a 'named graph' [2]. Despite the establishment of concepts such as NPs and named graphs, no unique way has been identified to facilitate the integration of all possible ways of publishing in a fast and reliable way. Here, we demonstrate how our interpretation of NPs, named graphs, knowledge resources and existing web tools can be combined to facilitate the integration of the diverse types of publishing and potentially lead to the discovery of new knowledge.

2 Practical Example of NPs: Feedback Looping

In this paper, we explore the generation and use of NPs [4] by means of a concrete example. We define a NP as a set of one or more RDF statements which assert some knowledge in the field of expertise of the person publishing, or demonstrate an endorsement of a statement by the latter. While the RDF annotations may be automatically obtained using text mining methods, they will have been manually approved and collected in a set, constituting a new object, by the author of the NP. In addition, it is important that a NP can be properly cited and its origin (provenance) traced. Therefore, a NP should include an identifier, e.g, digital object identifier (DOI), while provenance should include the author of the NP, and the origin of each of the statements included therein. Based on the above, we propose a model for NPs, which is summarised in Figure 1.

The model is cyclic and based on dynamic interaction between users and the machine. It consists of three steps, focusing on benefits for creators and users of NPs. Firstly, researchers create the NP and offer it to the broader community (see Figure 2). The second step involves machine consumption of the data generated by users. Machines can integrate data from multiple sources as long as they are represented in a common format with explicit semantics (RDF). For example, statements generated by a user can be integrated with statements coming from

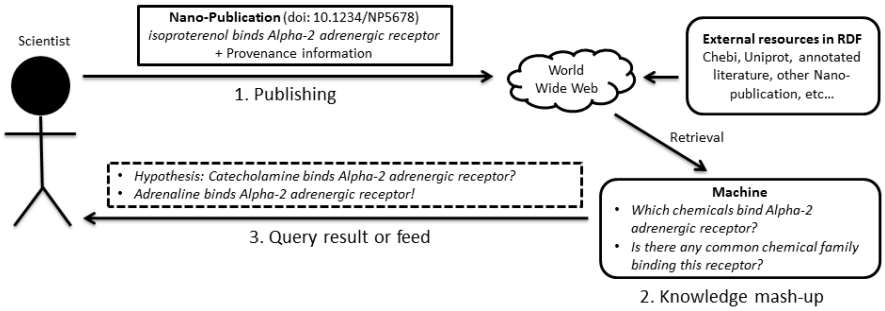

Fig. 1. A model for nano-publishing.

UniprotKB[7] or existing literature. Computers can also combine related NPs and generate new hypotheses. The third step of the model is the user's reward: feeds generated from the data integration done by the machine in the second step. The author of the NP will receive relevant information, based on previously published assertions processed by the machine. The hypotheses can be evaluated, rejected or validated by data added by the user, leading to a new NP, as described in step one. The dynamic human-computer interaction allows NP writers to access relevant information tailored to individualized retrieval, which they can enrich for the benefit of the community. As NPs are uniquely identified by a DOI, they can be cited and further used in any type of publication.

3 Working Example of NPs Using Semantic Wikis

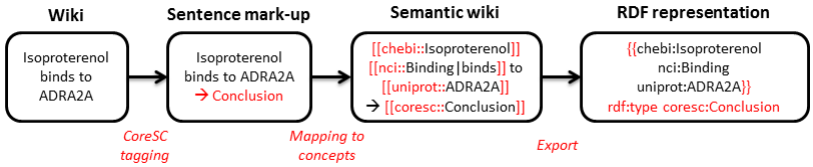

Fig. 2. General example of NP creation from a scientific statement contained in a wiki.

We propose that the route to a scientific NP can be facilitated by enabling the annotation of scientific notes or blogs at multiple levels of detail. The author of the notes can then package together aspects of their notes as a scientific publication. We have created a prototype of a tool for the open source wiki MediaWiki, allowing the user to manually annotate a scientific document with automated markup; multiple tags are allowed at the sentence level. This tool is aimed at scientists who use MediaWiki as an electronic lab notebook environment. Scientists could post a set of their annotated sentences as a NP. Additional annotation (manual or automated) of important entities and relations between entities can be provided for terms within these sentences. These can then be saved as triples linked to the NP. A mockup screenshot of entity-level markup can be found in Figure 3. The prototype of our tool can also model the scientific

[7] http://www.uniprot.org/help/uniprotkb

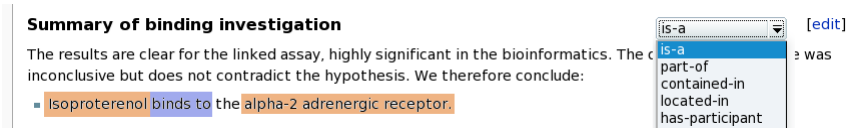

Fig. 3. Mock-up of annotation tool (automatic or manual) of significant entities and relations in a wiki used as a laboratory notebook.

discourse of the document or notes in terms of CoreSC [5], an annotation scheme successfully used to automatically recognise core scientific concepts in research articles. Thus we can retrieve the semantic context from which a NP has been generated (Result, Conclusion, Hypothesis, etc.). Once generated, the RDF form of the NP can then be exported and hosted on an external RDF hosting site. Figure 2 demonstrates this process.

The following example illustrates the format of NPs including a simple statement (Figure 4)[8]. The illustration is extended to show how a new hypothesis can be generated from the user's input.

```
@prefix chebi: <http://purl.org/obo/owl/CHEBI#> .
@prefix nci: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#> .
@prefix uniprot: <http://purl.uniprot.org/uniprot/> .
@prefix prv: <http://purl.org/net/provenance/ns#> .
@prefix coreSC: <http://www.sapientaproject.com/publications/coresc#>
{ chebi:CHEBI_6257 nci:C82888 uniprot:P08913 . }
  rdf:type prv:DataItem, coreSC:conclusion ;
  prv:createdBy [ rdf:type prv:DataCreation ;
    prv:performedAt "2009-07-10T12:00:00Z"^^xsd:dateTime ;
    prv:performedBy [a foaf:Person; foaf:name "Ben Huxley"] ;
    prv:accessedResource <http://www.mysite.com/mywiki/mydoc48#sentence65> ] ;
  _:uri doi:10.1234/NP5678 .
```

Fig. 4. RDF representation of exemplary conclusion.

A researcher derives the conclusion: "isoproterenol binds to the Alpha-2 adrenergic receptor" and decides to create a NP corresponding to the conclusion. In the semantically enriched statement, *isoproterenol* is mapped to the ChEBI ontology (ChEBI)[9], the receptor to UniprotKB, and the action *binds* to one of the terms of the NCI Thesaurus[10]. Automated URI mappings can be achieved with services like the NCBO BioPortal[11]. Once the data has been transformed to RDF, the NP is available for further processing and integration into the network of linked data.

```
select ?chemical where {
  ?chemical rdf:type nci:C48807 .
  ?chemical nci:C82888 uniprot:P08913 .
}
```

Fig. 5. Query statement.

The example in Figure 5 illustrates how a generic query can be generated from the NP presented above. The query aims to retrieve other chemicals binding the receptor and is run over a knowledge base. The result is a list of known chemicals binding P08913. The adrenaline molecule (chebi:CHEBI_33568) will appear in the list and could be reported to the user. Applying a reasoner will reveal that adrenaline and *isoproterenol* are both members of the catecholamine family (chebi:CHEBI_33567). From this observation and in absence of any other information present in the databases, the following

[8] For provenance terminology and concepts we follow: http://trdf.sourceforge.net/provenance/ns.html
[9] http://www.ebi.ac.uk/chebi/
[10] http://ncit.nci.nih.gov/
[11] http://bioportal.bioontology.org/

hypothesis can be sent to the creator of the NP: "Do compounds from the catecholamine family bind the P08913 receptor?". The newly formed hypothesis can give rise to new experiments and be published in return (Figure 6), citing both the NP above and ChEBI, which are the basis of the hypothesis.

```
{ CHEBI:33567 nci:C82888 uniprot:P08913 } rdf:type prv:DataItem ;
  prv:createdBy [ rdf:type prv:DataCreation ;
    prv:performedAt "2009-07-10T12:00:00Z"^^xsd:dateTime ;
    prv:performedBy [a foaf:Person; foaf:name "Dan Brickley"] ;
    prv:usedData _:otherNanopub , _:chebiWeb ] ;
  _:uri doi:10.1234/OQ6789 .

_:otherNanopub rdf:type prv:DataItem ;
  prv:retrievedBy [ rdf:type prv:DataAccess ; prv:accessedResource doi:10.1234/NP5678 ] .
_:chebiWeb rdf:type prv:DataItem ;
  prv:retrievedBy [ rdf:type prv:DataAccess ;
  prv:accessedResource <http://www.ebi.ac.uk/chebi/displayAutoXrefs.do?chebiId=CHEBI:33568> ] .
```

Fig. 6. Potentially new NP.

4 Conclusion

We have described an infrastructure for creating NPs using web tools such as blogs and wikis, while integrating information from a number of external resources. We also demonstrated how the produced NPs can be extended via the integration of information from other knowledge resources through querying across available resources on a semantic layer. A researcher using this system can decide whether or not to confirm the result of a query resulting from their initial publication and publish it as a new NP. Enabling this strategy of publishing will not only facilitate the integration of diverse resources but also allow for fast and precise knowledge dissemination and retrieval.

References

1. Attwood, T.K., et al.: Utopia documents: linking scholarly literature with research data. Bioinformatics 26(18), i568–74 (2010)
2. Carroll, J.J., et al.: Named graphs, provenance and trust. International World Wide Web Conference (2005)
3. Gerstein, M., et al.: Structured digital abstract makes text mining easy. Nature 447(7141), 142 (2007)
4. Groth, P., Gibson, A., Stickler, P.: The anatomy of a nanopublication. Information Services and Use 30(1), 51–56 (2010)
5. Liakata, M., et al.: Corpora for the conceptualisation and zoning of scientific papers. In: International Conference on Language Resources and Evaluation (2010)
6. Mons, B., Velterop, J.: Nano-publication in the e-Science era. Workshop on Semantic Web Applications in Scientific Discourse (Jan 2009)
7. Passant, A., Ciccarese, P., Breslin, J., Clark, T.: SWAN/SIOC: aligning scientific discourse representation and social semantics. Workshop on Semantic Web Applications in Scientific Discourse (2009)