

## Article (refereed) - postprint

---

Wright, Daniel G.; Harrison, Kathryn A.; Watkins, John. 2015. **Automated tagging of environmental data using a novel SKOS formatted environmental thesaurus** [in special issue: Semantic e-sciences] *Earth Science Informatics*, 8 (1). 103-110. [10.1007/s12145-014-0183-1](https://doi.org/10.1007/s12145-014-0183-1)

© Springer-Verlag Berlin Heidelberg 2014

This version available <http://nora.nerc.ac.uk/508638/>

NERC has developed NORA to enable users to access research outputs wholly or partially funded by NERC. Copyright and other rights for material on this site are retained by the rights owners. Users should read the terms and conditions of use of this material at <http://nora.nerc.ac.uk/policies.html#access>

**This document is the author's final manuscript version of the journal article, incorporating any revisions agreed during the peer review process. Some differences between this and the publisher's version remain. You are advised to consult the publisher's version if you wish to cite from this article.**

**The final publication is available at Springer via <http://dx.doi.org/10.1007/s12145-014-0183-1>**

Contact CEH NORA team at  
[noraceh@ceh.ac.uk](mailto:noraceh@ceh.ac.uk)

1

2 **Title:** Automated tagging of environmental data using a novel SKOS formatted  
3 environmental thesaurus.

4

5 **Authors:** Daniel G. Wright, Kathryn A. Harrison, John Watkins

6

7

8 **Affiliation:** Centre for Ecology & Hydrology, Lancaster Environment Centre, Library Avenue,  
9 Bailrigg, Lancaster, LA1 4AP, UK

10

11 **Corresponding author:** Daniel G, Wright

12 **Email:** dgwr@ceh.ac.uk

13

14

15

16

17

18 **Abstract**

19 There is increasing need to use the widest range of data to address issues of environmental  
20 management and change, which is reflected in increasing emphasis from government  
21 funding agencies for better management and access to environmental data. Bringing  
22 together different environmental datasets to confidently enable integrated analysis requires  
23 reference to common standards and definitions, which are frequently lacking in  
24 environmental data, due to the broad subject area and lack of metadata. Automatic  
25 inclusion within datasets of controlled vocabulary concepts from publicly available standard  
26 vocabularies facilitates accurate annotation and promotes efficiency of metadata creation.  
27 To this end, we have developed a thesaurus capable of describing environmental chemistry  
28 datasets. We demonstrate a novel method for tagging datasets, via insertion of this  
29 thesaurus into a Laboratory Information Management System, enabling automated tagging  
30 of data, thus promoting semantic interoperability between tagged data resources. Being web  
31 available, and formatted using the Simple Knowledge Organisation System (SKOS)  
32 semantic standard, this thesaurus is capable of providing links both to and from other  
33 relevant thesauri, thus facilitating a linked data approach. Future developments will see  
34 extension of the thesaurus by the user community, in terms of both concepts included and  
35 links to externally hosted vocabularies. By employing a Linked Open Data approach, we  
36 anticipate that Web-based tools will be able to use concepts from the thesaurus to discover  
37 and link data to other information sources, including use in national assessment of the extent  
38 and condition of environmental resources.

39

40

41

42 **Keywords:** Thesaurus, vocabularies, metadata, annotation, standards, SKOS, data tagging,  
43 environmental chemistry

## 44 **Introduction**

45 Disparate vocabularies and sparse descriptions present in environmental data are an  
46 impediment to gaining greatest value from these data when considering their re-use or  
47 integration (Michener et al. 1997). The concept of Linked Open Data (LOD), first proposed  
48 by Tim Berners-Lee, refers to a set of best practices for publishing and connecting structured  
49 data on the web. The concept of standardised web-accessible links within data, or  
50 documents, can be used to address issues of interoperability, within the field of  
51 environmental data, as described here, or any other discipline. The creation of a 'world wide  
52 web of data' whereby pieces of data and information are semantically related to other  
53 relevant information can greatly enhance the user's ability to derive additional value about a  
54 concept with little extra effort (Bizer et al. 2009), and can facilitate interoperability both within  
55 and across domains. Here, we describe the preliminary steps to implementing these  
56 concepts within a national environmental chemistry analysis facility, funded by the UK  
57 Natural Environment Research Council (NERC), though the steps and issues discussed will  
58 be relevant to any area of research wishing to promote interoperability and re-use in their  
59 sector.

60

61 There is increasing pressure on publicly funded research institutes to demonstrate value for  
62 money in the data they produce and enable others to re-use and add value to these data  
63 (e.g. Research Councils UK (RCUK) common principles on data policy). However, data  
64 lacking description of the methodologies used and/or measurements collected hinders this  
65 process (LeDuc et al. 2007). Inadequate annotation of data has been particularly prevalent  
66 within the field of ecology, where documentation of this information is often lacking or an  
67 afterthought in many projects (Madin et al. 2007). The ability to generate metadata during  
68 the creation of data in a pre-defined way would save inefficient use of scarce staffing  
69 resources for manual documentation work (Batcheller 2008). Using vocabularies in

70 retrospective creation of metadata (for example, after a project or period of work has come  
71 to an end), can be an especially time consuming process with increased risk of errors  
72 occurring in metadata and vocabulary tags. Further, early selection of vocabularies intended  
73 for use in a project will enable identification of any missing concepts, for which there is no  
74 acceptable existing vocabulary term. This allows adequate time to contact relevant  
75 vocabulary governance groups to request new concepts, rather than attempting to add new  
76 concepts to existing vocabularies at the end of a project, when timescales for completion of  
77 work are often compressed, and may be insufficient to allow for requests to external  
78 agencies to be processed. Once the desired vocabularies have been selected, development  
79 of automated methods for tagging datasets provides the advantages of minimising the time  
80 required for tagging and increasing the accuracy with which it is carried out, since it reduces  
81 human error (Ailamaki et al. 2010). Data can also be tagged at point of source i.e. as it is  
82 produced, further reducing the likelihood of errors occurring. Deployment of such  
83 methodologies to aid in automated tagging of datasets with required information for re-use  
84 and integration will be of great benefit to the environmental and ecological communities, but  
85 will also ensure that the resources produced are able to be re-purposed by any community  
86 wishing to make use of them, without recourse to the original data generators. It is also  
87 apparent that many data generators do not fully realise the benefits of using vocabularies to  
88 describe the data they produce, and do not therefore utilise vocabularies at all, thus  
89 devaluing the datasets produced. By automating the process of tagging using vocabulary  
90 concepts, the onus to employ vocabularies is removed from the data generator, which will  
91 hopefully increase the use of vocabularies within the research community.

92

93 The location, integration and re-use of data have particular scientific value when running  
94 meta-analyses which can be a very powerful way of answering complex multi-disciplinary  
95 questions (Treseder 2004). However, this method has been criticised historically for not  
96 comparing like with like (Arnqvist and Wooster 1995). Automated tagging can ensure

97 disparate datasets are semantically comparable and therefore potentially interoperable  
98 through the use of Web-accessible controlled vocabularies. Integration between datasets  
99 tagged using concepts from the same Web-accessible vocabulary, or described using a  
100 vocabulary which has public mappings to another Web-accessible vocabulary, is  
101 considerably easier than integrating datasets which do not utilise a vocabulary or are  
102 described using a separate vocabulary to which no mappings exist. This can be particularly  
103 important where data are being employed for novel purposes, not originally considered by  
104 the initial project responsible for generating the data, or in attempting to utilise data from a  
105 different discipline e.g. atmospheric scientists wishing to make use of oceanographic data,  
106 etc. Discovery of relevant datasets is also facilitated if keywords provided in discovery  
107 metadata are selected from defined vocabularies so Web search engines can identify  
108 datasets from Web-enabled data catalogues. Use of Web-accessible vocabularies can also  
109 reduce the amount of content-level or contextual metadata which must be provided  
110 alongside a dataset to permit its re-use, as all the definitions and supplementary information,  
111 including semantic relations, on the concepts can easily be read from the Web. This practise  
112 has the two-fold benefit of saving data producers' time by reducing the amount of metadata  
113 that they are required to produce, and also reducing the volume of metadata required to be  
114 processed and maintained per dataset.

115 The Centre for Ecology and Hydrology (CEH) operates a centralised analytical chemistry  
116 facility processing samples from NERC funded researchers and the long-term monitoring  
117 activities within CEH. Implementation of a single appropriate vocabulary within this analytical  
118 chemistry facility would have the potential to improve re-use, interoperability and discovery  
119 of all datasets produced via this laboratory for a wide range of researchers. This is of  
120 particular importance, given new drives to make data open and freely available where it has  
121 been publicly funded, where the original data generators often do not have knowledge of  
122 who has accessed the dataset, thus making unambiguous description of the data essential.

123 The introduction of a new Laboratory Information Management System (LIMS) within the  
124 Analytical Chemistry Group provided the ideal opportunity to experiment with the idea of  
125 automatic tagging of data at source using a controlled vocabulary. A Laboratory Information  
126 System is used to control and manage samples, standards, test results, reports, laboratory  
127 staff, instruments, and work flow automation (Skobelev et al. 2011). If an appropriate  
128 vocabulary could be inserted into the LIMS, then any dataset produced would automatically  
129 be tagged with concepts from that thesaurus. This would provide the foundation for not only  
130 linking the broad range of datasets produced through this facility, but also linking these data  
131 with data produced elsewhere that contain comparable vocabulary tags. This foundation  
132 could then be used for future development of Linked Open Data where these data can be  
133 made discoverable and accessible by incorporating concepts from vocabularies into Web  
134 search and delivery tools.

135

### 136 **Thesaurus creation**

137 To implement automated semantic tagging with the new LIMS required identification or  
138 development of a suitable vocabulary. Rather than simply describing the determinands being  
139 measured, it was also necessary that any vocabulary would include concepts covering units  
140 of measurement, analytical methods and types of machine/instrument used. We also  
141 required that any vocabulary employed would be freely available to the public as an online  
142 resource, with concepts identified by a uniform resource identifier (URI), which would be  
143 beneficial for a number of reasons. First, this would promote use of the vocabulary by  
144 allowing external users to access and use concepts from the vocabulary for description of  
145 their own datasets. Second, it would also facilitate re-use of any dataset tagged using the  
146 concept, as all required information about a concept can easily be obtained simply by  
147 entering the URI into a web browser, meaning it would not have to be provided alongside the  
148 data as contextual metadata. Third, and perhaps most importantly, by being available in this

149 manner, it would also permit mappings to other online vocabularies using linked data  
150 approaches, thus greatly enhancing the volume of information about a concept that is  
151 available to users. The concepts contained in the vocabulary could be linked, directly or  
152 indirectly, to concepts from vocabularies developed for use in other domains, thus increasing  
153 the number of data resources which can be integrated with datasets tagged using the  
154 original vocabulary, and not limiting their use to within their original scientific domain.

155  
156 Latre et al. (2012) suggest that there are four common steps in the process of thesaurus  
157 creation, though each of these steps can be approached in a variety of ways. The first step  
158 is to review other available thesauri – it is better to re-use an existing thesaurus that is fit for  
159 purpose and potentially already has a user community, than to automatically create a new  
160 thesaurus, which would lead to a proliferation of redundant thesauri. Second, developers of  
161 thesauri need to decide how they wish to structure the thesaurus, and how it will be  
162 formatted. Third, the candidate terms for inclusion in the thesaurus must be selected before  
163 undergoing the final step, where the potential concepts are reviewed and validated against  
164 the agreed standard. The approaches we employed for each of these steps are outlined in  
165 the following sections.

166

## 167 **1. Reviewing existing thesauri**

168 Although several existing vocabularies (e.g. Chemical Entities of Biological Interest<sup>1</sup> (ChEBI)  
169 and Chemical Methods Ontology<sup>2</sup> (CMO)) fulfilled one or more of the required criteria, no  
170 single candidate vocabulary contained all the categories of concepts that we wished to  
171 include or described concepts with the required level of detail. Similarly, following a  
172 preliminary inspection of legacy datasets held by CEH, it became apparent that many of the  
173 determinands measured did not fit well in established ontologies/vocabularies such as

---

<sup>1</sup> <http://www.ebi.ac.uk/chebi/>

<sup>2</sup> <http://www.rsc.org/ontologies/CMO/>



174 ChEBI, consisting as they did of several different forms of elements grouped together due to  
175 their method of analysis. Although Simons et al. (2013) describe a method by which this  
176 could be accommodated, by employing the Observable Properties Model to completely  
177 separate substances from quantity/kind and units, early discussions with the analytical  
178 chemistry team regarding the LIMS revealed that we would have a limited number of fields  
179 available to describe the whole analytical process for each analyte, so a certain degree of  
180 concatenation between quantity/kind and the substance was required. Further, many of the  
181 units to be described were non-standard, or derived units, which were specifically related to  
182 particular determinands or methods (e.g. microsiemens per centimetre at 25 degrees  
183 Celsius), or were not derived from those already contained in existing vocabularies, such as  
184 the Quantities, Units, Dimensions and Data Types<sup>3</sup> (QUDT) ontologies (e.g. micrograms per  
185 gram (dry weight)) . We also wished to retain control over the governance of any developed  
186 vocabulary, so that updates could be made quickly, as required, and new concepts that were  
187 essential to our user community could be added without requiring approval from external  
188 governors. Another point for consideration was that only limited time to assist in vocabulary  
189 development was available from domain experts. It was felt that this was best employed by  
190 obtaining labels and definitions for concepts from the experts, rather than asking them to  
191 consider extensive lists of candidate concepts, and make a decision as to whether any of  
192 them met our requirements, or not, in which case more time would have to be spent  
193 identifying alternative candidate concepts for consideration. Consequently it was decided  
194 that we would develop a new vocabulary – the CEH Analytical Services Thesarasus (CAST).

195

## 196 **2. Modelling the thesaurus**

197

### 198 *2.1 Organisation System*

---

<sup>3</sup> <http://qudt.org/>

199 An initial step was to decide how the thesaurus would be structured. Given that the  
200 vocabulary would need to include concepts covering different areas of the analytical  
201 process, it was desirable that any proposed way of structuring the vocabulary could  
202 accommodate the requirement to split concepts into clearly defined groupings or facets. The  
203 structure should also have the ability to describe relationships between the concepts  
204 selected for inclusion, something a flat list of defined terms would not achieve. Some sort of  
205 knowledge organisation system (KOS) was required in order to structure the thesaurus.  
206 There were two obvious candidates for this: the Web Ontology Language (OWL) (World  
207 Wide Web Consortium 2012) and the Simple Knowledge Organisation System (SKOS)  
208 (World Wide Web Consortium 2009). OWL would be a heavyweight option, but a  
209 semantically rich one, capable of expressing any number of desired relationships between  
210 classes and individuals, whereas SKOS is a much more lightweight approach, with a more  
211 limited set of properties for describing relationships between concepts. OWL has been used  
212 very effectively in modelling ontologies, such as Chemical Entities of Biological Interest  
213 (ChEBI), but would require a large degree of input from domain specialists in order to agree  
214 the nature of the semantic relationships to be deployed. SKOS is a formal language for  
215 representing controlled structured vocabularies, including thesauri, classification schemes,  
216 taxonomies and subject heading systems (Miles and Pérez-Agüera 2007) as well as being a  
217 World Wide Web Consortium (W3C) recommendation for providing a standard way of  
218 organising knowledge using the Resource Description Framework (RDF) (World Wide Web  
219 Consortium 2004). Given that the primary objective for the development was to provide a  
220 simple reference system for use by scientists, SKOS was more suitable to our needs. It was  
221 therefore decided that CAST would be created using SKOS. The benefits of using SKOS  
222 would be that hierarchical, and other relationships between concepts could be easily  
223 represented using the suite of relationships defined in the SKOS standard (e.g. broader,  
224 narrower and related). In addition, it would also enable mappings between CAST and other  
225 selected vocabularies, such as ChEBI, thus allowing integration between datasets tagged  
226 using concepts from CAST and other vocabularies to which CAST had been mapped.

227

## 228 *2.2 SKOS Editor*

229 In order to create a SKOS formatted thesaurus, an editing tool was required. To allow the  
230 CAST to be publicly available, a method of accessing the thesaurus over the web was also  
231 necessary. We selected the commercial application ‘PoolParty<sup>4</sup>’ for creation and hosting of  
232 CAST as it allowed us to fulfil the above stated criteria. Other options were available, but  
233 were rejected due to not having all the required functionality or being prohibitively expensive.  
234 PoolParty permits users to create and edit SKOS formatted vocabularies, supporting linked  
235 data approaches via mappings to other resources in the Linking Open Data (LOD) Cloud  
236 and other vocabularies hosted within PoolParty. This was important given our future desire  
237 to define mappings between CAST and other selected vocabularies. Of particular interest  
238 was the previously mentioned Chemical Entities of Biological Interest (ChEBI), which is one  
239 of the ontologies of the Open Biological and Biomedical Ontologies<sup>5</sup> (OBO) Foundry. The  
240 benefits of forging these links would be that as ChEBI is an extremely rich vocabulary, it  
241 provides more information on the concepts it contains than CAST, as it is structured using  
242 OWL rather than SKOS. However, mappings between CAST and ChEBI could be achieved  
243 by using relationships defined by SKOS for linking to external resources, simply by utilising  
244 the URIs of concepts contained in ChEBI. Thus, by mapping between CAST and ChEBI  
245 concepts, we can add CAST into a linked data network, allowing users to access a wealth of  
246 additional information relating to concepts in CAST than would otherwise be available. This  
247 has the added benefit of facilitating integration between datasets tagged using concepts  
248 from both CAST and ChEBI. Importantly, PoolParty also keeps track of alterations, via  
249 changes to individual concepts, creating an audit trail of edits made to the thesaurus. Use of  
250 PoolParty also facilitated development of the thesaurus in private via restriction of access to  
251 a group of developers, prior to making it publicly available once the initial batch of concepts  
252 had been created and defined.

---

<sup>4</sup> <http://www.poolparty.biz/>

<sup>5</sup> <http://www.obofoundry.org/>

253

254 **3. Concept selection**

255 There are two methods of selecting concepts for inclusion; top-down, where the groupings  
256 into which concepts will fall are defined, or bottom-up, where all the concepts requiring  
257 description are identified and natural groupings are subsequently defined (Latre et al. 2012).  
258 In this instance it was decided to adopt a top-down approach to identify concepts for  
259 inclusion, given that areas of the analytical process requiring description already existed.  
260 The first steps in selecting concepts for inclusion in CAST involved identification of the facets  
261 required to cover the elements to be included in the vocabulary, such as determinands being  
262 measured and the processes involved in their measurement. SKOS permits two alternative  
263 options for modelling of these facets – they can either be as Top Concepts of a Concept  
264 Scheme (approximately equivalent to a standard vocabulary), where the Top Concepts  
265 represent the broadest level of the facet being represented, or by collecting concepts  
266 comprising each facet as Collections (World Wide Web Consortium 2004). Of the two, the  
267 best method to employ frequently depends on the application being used, and it is often  
268 more intuitive to deploy the first of these approaches where a navigation hierarchy is  
269 required (World Wide Web Consortium 2004). Given that the primary objective in developing  
270 the thesaurus was to provide a reference source to enable interoperability between datasets,  
271 it was decided that the most appropriate strategy would be to instantiate the required facets  
272 as Top Concepts for a Concept Scheme, using the property *topConceptOf*. Top Concepts  
273 were selected broadly corresponding to table and field names from a relational database  
274 schema which had previously been designed to store legacy hydrochemistry data. The  
275 database itself was never actually implemented, but it was felt that it provided a sound basis  
276 for identification of Top Concepts as it suggested areas of metadata which would be  
277 produced for any dataset created by the analytical chemistry facility. The basis for this was  
278 that any measurement would be of something (i.e. the thing being measured), which would  
279 have some kind of unit, and would also have been measured using some overall

280 methodology, which could primarily be described using a method of analysis. Secondly, a  
281 component of the overall methodology could include details of how samples had been  
282 preserved and filtered, in addition to the category and model of machine/instrument that had  
283 been used to perform the analysis, though these would not always be relevant for every  
284 analysis. Each concept in a facet could potentially be associated with many other concepts  
285 in other facets, thus producing associations between the different facets as illustrated in Fig.  
286 1, though these relationships would not be formalised semantically in the thesaurus. If  
287 domain experts subsequently desired inclusion of a new facet, it would be possible to add  
288 additional Top Concepts at a later date, to support this. Initial investigation provided the  
289 following Top Concepts requiring population, which were defined as follows:

290 Determinands – aspects of a sample or feature which are measured and assigned a value  
291 from an agreed domain

292 Measurement units – units used for measurement of determinands

293 Machine descriptions – descriptions of machines/instruments used for analyses

294 Methods – methods used for sample or feature analysis

295 Filtration – filtration methods applied to samples

296 Preservation – preservation methods applied to samples

297 Candidates for narrower concepts to each of these Top Concepts were selected from  
298 metadata for the legacy hydrochemistry dataset, with a preferred label, alternative label/s,  
299 definition and semantic relationships to other concepts, provided for each concept. Possible  
300 relationships included broader and narrower (hierarchical), and related (associative), as  
301 defined within SKOS. The majority of relationships defined within the thesaurus would be  
302 hierarchical e.g. acid recoverable boron is a broader concept than dissolved boron, as its  
303 definitions states it includes the dissolved fraction plus particulates dissolved by acidification.

304 Approval of concepts was achieved via an iterative process of sending concepts for

305 consideration by domain experts within the organisation, making amendments to concepts,  
306 and resubmission to the domain experts, until all parties were satisfied with the information  
307 available and defined relationships for each concept.

308 Further, a means for addition of new concepts to the thesaurus would need defining to  
309 accommodate measurement of new determinands, development of new methods and/or  
310 deployment of new machines by the analytical chemistry team. It was clear that the  
311 thesaurus could not be a static object – it would be a live one which would require  
312 maintenance in order to retain its relevance. Therefore, a decision was made to adopt the  
313 approach of populating CAST with a selection of concepts describing the most frequently  
314 used analyses and releasing it to coincide with the implementation of a new LIMS in  
315 analytical chemistry. Once the initial selection of concepts had been approved and created in  
316 CAST, the status of the thesaurus within PoolParty could be altered to 'public', meaning that  
317 it would be freely accessible to all at <http://onto.nerc.ac.uk/CAST>. New concepts could then  
318 be added to CAST, as the need arose, via the mechanism detailed below.

319

#### 320 **4. Reviewing selected concepts against the standard**

321 The thesaurus was developed according to the American National Standards Institute  
322 standard for development of monolingual controlled vocabularies (National Information  
323 Standards Organization 2005) which is a freely available and recognised standard in this  
324 discipline, proven via development of vocabularies across many domains (Latre et al.  
325 2012). The standard provided a specification for the grammatical form of preferred labels for  
326 concepts and methods for selecting the preferred form, such as selecting the mostly  
327 commonly used lexical variants, within the scientific community, for concepts, and avoiding  
328 the use of upper-case letters except in the case of proper nouns. Lexical variants not  
329 selected, or abbreviations, were included as alternative labels for concepts. Once preferred

330 labels and definitions had been agreed with domain experts, they were checked for  
331 conformance against the standard, and amended if necessary.

332

### 333 **User interface development**

334 To enable users to access the relevant information for each concept a web accessible user  
335 interface was required. This interface needed to be human readable, clearly displaying  
336 labels for the concepts, the definition and any relationships to other concepts (both internal  
337 and external to the thesaurus). PoolParty provides a basic template for a user interface, but  
338 it was not suitable in its current format. Therefore, the template was modified significantly by  
339 our own developers, in order to display the information required in a clear and accessible  
340 manner. Once this was in place, users could take a URI for any concept, enter it into a  
341 browser and immediately land on a page containing all the information about that concept  
342 (Fig. 2).

343

344

### 345 **Governance**

346 Once the thesaurus had been made publicly available, a mechanism to allow for addition of  
347 new concepts, identified either by laboratory managers or by users planning to produce a  
348 dataset containing determinands, methods or units not already contained in the thesaurus,  
349 was clearly required. To this end, an email account linked to a task-tracking system was  
350 created which allowed users to suggest new concepts they would like included in CAST,  
351 including a proposed preferred label and definition for the concept. This account would  
352 initially be checked by the CAST gatekeepers against other entries in the thesaurus to avoid  
353 duplication, and against the relevant standard to ensure compliance, before being passed on  
354 to the CAST Governance Group (CGG), a panel of domain experts, who would decide on  
355 the suitability of the concept for inclusion, define any relationships to other concepts in the

356 thesaurus and make any required changes to the suggested preferred label and definition.  
357 These concepts would then be passed back to the CAST gatekeepers who would insert the  
358 accepted concepts into the thesaurus, who would subsequently notify the laboratory  
359 managers in order that the new concepts could be added in to the LIMS.

360

### 361 **Deployment in the Laboratory Information Management System**

362 The primary objective in developing CAST was that it could be inserted, by manually  
363 inputting Uniform Resource Identifiers (URIs) for individual concepts, directly into the LIMS  
364 used by CEH Analytical Services. Once the URIs for the concepts had been entered into the  
365 LIMS, and associated with the correct determinands, methods and units, no further human  
366 input was required in order to produce tagged datasets, other than setting the LIMS to  
367 perform the required analyses. The LIMS would then analyse the samples, as programmed,  
368 but in addition to outputting the results, it would also include the URIs from concepts from  
369 CAST, for the analyses it had performed, in the output file it produced. This output takes the  
370 form of a comma separated value (csv) file of results, with the URIs (e.g.  
371 <http://onto.nerc.ac.uk/CAST/13>) being present in columns alongside columns containing  
372 human readable labels (e.g. dissolved ammonium) for the relevant determinands, units and  
373 methods. This automatic tagging removes the requirement for researchers to spend time  
374 manually tagging their dataset using concepts from a vocabulary, and ensures that all  
375 datasets produced by the facility are tagged using the same vocabulary, increasing their  
376 potential re-use value and allowing integration between tagged datasets, in addition to  
377 potentially providing a wealth of additional information to users simply by dereferencing the  
378 URIs.

379

### 380 **Future developments**



381 Long-term, the objective is that CAST will provide a comprehensive thesaurus containing  
382 concepts capable of describing determinands, units, analytical methods and machines used  
383 within environmental chemistry research that is publicly accessible for use in tagging data or  
384 linking to other related LOD standard vocabularies. To achieve this will require active  
385 participation from users in order to both continually improve and expand the thesaurus, and  
386 create the links to external resources. These will include, where possible, links to  
387 vocabularies for units of measurement, such as QUDT, and also to ChEBI. These will be  
388 specified using the standard SKOS relationships for linking to external resources of  
389 *exactMatch*, *broadMatch* or *narrowMatch* where appropriate. Whilst this will require  
390 significant input from domain experts, the benefits to be gained by increased interoperability  
391 make this an obvious area for further investment.

392

393 The automated tagging of datasets, such as that performed by the LIMS, is extremely  
394 efficient, given that manually tagging datasets is a time-consuming and expensive process  
395 (Batcheller 2008), and it also allows laboratory managers to quickly and easily identify gaps  
396 in the thesaurus to be filled, as there will be determinands/units/methods which do not have  
397 an associated concept URI in the csv outputs, which can be identified by laboratory  
398 managers when inspecting the output files. It also removes the opportunity for dataset  
399 authors to make an error when tagging their dataset, as it is received from the Analytical  
400 Chemistry facility already containing URI tags for every concept contained in the dataset.  
401 Development of CAST means that the measurements made by this analytical chemistry  
402 facility are now identified by URIs and support LOD approaches to data management. One  
403 such approach of interest is the ability to link chemical measurements to the location from  
404 which they were sampled (e.g. field site). In turn, this means that chemical measurements  
405 made at a site can also be linked to biodiversity and habitat data collected from the same  
406 location. This linking would enable quick and easy querying of previously disparate datasets  
407 e.g. determining the chemical composition of the habitat associated with plant

408 species/functional groups. This work is being undertaken to support research into extent and  
409 state of natural capital assets (e.g. Woodland, Soils, and Biodiversity) and which data sets  
410 can be used to quantify them.

411

412 Further, building on the idea of using CAST to link important environmental monitoring data  
413 sets to national ecosystem and natural capital assessment; we will be working with sister  
414 research institutes such as the British Geological Survey and the British Oceanographic  
415 Data Centre in order to link concepts used in their environmental monitoring programmes to  
416 the work carried out in developing CAST. We hope that this will enable future Web searches  
417 to identify a wide range of data relating to the particular environmental concepts and enable  
418 them to be integrated with confidence using the standardised description of measurements  
419 and methods that are easily accessible via automatically generated Web links.

420

## 421 **Conclusions**

422 Implementation of the approaches described here has enabled accurate semantic  
423 interoperability between environmental chemistry datasets tagged using CAST, which has  
424 proven invaluable in a current project which aims to link environmental data from across  
425 NERC with the intention of being able to quickly assess where, spatially, analytes have been  
426 measured, regardless of the individual project or organisation responsible for collection of  
427 data. This has only been possible through use of a common vocabulary, which has been  
428 mapped to other discipline specific vocabularies. Use of CAST has also promoted re-use of  
429 data; well-defined datasets are easier for researcher to subsequently re-use as they are able  
430 to quickly understand what has been measured and how data has been generated. Further,  
431 Data Centres, such as NERC's Environmental Information Data Centre (EIDC) are  
432 increasingly requiring depositors of data to provide more detailed supporting information for  
433 datasets – material which can easily be provided using a web-accessible, publicly available

434 vocabulary to describe data. Automation of the tagging process, via a laboratory information  
435 management system, has increased efficiency of metadata authoring and reduced the  
436 likelihood of errors occurring. By using semantic standards for development of CAST, we  
437 have ensured that the thesaurus is fully compatible with Linked Open Data standards. Future  
438 developments will see extensions to CAST by the user community, in terms of both concepts  
439 included and links to externally hosted vocabularies enabling links to a wide range of publicly  
440 funded environmental data. Through use of a Linked Open Data approach, we anticipate  
441 that Web-based tools will be able to use CAST concepts to discover and link data to other  
442 information sources, including use in national assessment of the extent and condition of  
443 environmental resources.

444

#### 445 **Acknowledgements**

446 Thanks to Nic Bertrand for comments on a previous manuscript, and to Darren Sleep, Colin  
447 Neal, Phil Rowland and Gloria Dos Santos Pereira for assistance in development of CAST.  
448 Thanks also to Chris Johnson and Sabera Patel for developing the user interface.

449

#### 450 **References**

- 451 Ailamaki, A, Kantere V, Dash D (2010) Managing scientific data. *Commun. ACM* 53:68-78
- 452 Arnqvist G, Wooster D (1995) Meta-analysis: Synthesizing research findings in ecology and  
453 evolution. *Trends in Ecology and Evolution* 10: 236-240.
- 454 Batcheller JK (2008) Automating geospatial metadata generation – An integrated data  
455 management and documentation approach. *Comput. Geosci.* 34:387-398
- 456 Bizer C, Heath T, Berners-Lee T (2009) Linked data – The story so far. *Semantic Web and*  
457 *Int. J. Semantic Web Inf. Syst.* 5:1-22

458 Latre MA, Hofer B, Lacasta J, Nogueras-Iso J (2012) The Development and Interlinkage of  
459 a Drought Vocabulary in the EuroGEOSS Interoperable Catalogue Infrastructure.  
460 International Journal of Spatial Data Infrastructures Research. 7:225-248.

461 Le Duc MG, Yang L, Marrs RH (2007) A database application for long-term ecological field  
462 experiments. Journal of Vegetation Science 18: 509-516.

463 Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D, Villa F (2007) An ontology for  
464 describing and synthesizing ecological observation data. Ecological Informatics 2: 279-296.

465 Michener, WK, James WB, Helly J, Kirchner TB, Stafford SG (1997) Non-geospatial  
466 metadata for the ecological sciences. Ecological Applications 7:330-342.

467 Miles A, Pérez-Agüera JR (2007) SKOS: Simple Knowledge Organisation  
468 for the Web. Cataloging and Classification Quarterly, 43(3-4): 69-83.

469 National Information Standards Organization. (2005) Guidelines for the Construction,  
470 Format, and Management of Monolingual Controlled Vocabularies. ANSI/NISO Z39.19-2005

471 RCUK (2014) RCUK Common principles on data policy  
472 <http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx> Accessed 08 January 2014.

473 Simons, BA, Yu, J, Cox, SJD (2013) Defining a water quality vocabulary using QUDT and  
474 ChEBI. In: Piantadosi, J, Anderssen, RS, Boland, J (Eds.) MODSIM2013, 20th International  
475 Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and  
476 New Zealand, Adelaide: 2548-2554.

477 Skobelev, DO, Zaytseva, TM, Kozlov AD, Perepelitsa VL, Makarova AS (2011) Laboratory  
478 information management systems in the work of the analytic laboratory. Measurement  
479 Techniques 53:1182-1189.

480 Treseder KK (2004) A meta-analysis of mycorrhizal responses to nitrogen, phosphorous,  
481 and atmospheric CO<sub>2</sub> in field studies. New Phytologist 164:347-355. World Wide Web

482 Consortium (2004) RDF Primer. <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>  
483 Accessed 08 January 2014.

484 World Wide Web Consortium (2009) SKOS Simple Knowledge Organization System Primer.  
485 <http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/> Accessed 08 January 2014.

486 World Wide Web Consortium (2012) OWL 2 Web Ontology Language (Second Edition)  
487 <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/> Accessed 21 July 2014.

488

### 489 **Figure captions**

490

491 **Fig. 1** Graph showing initial ideas of how facets could be related within the thesaurus

492

493 **Fig. 2** User interface for CAST, showing the preferred label, alternative labels, URI, definition  
494 and relationships for the selected concept

495

496

Figure 1.

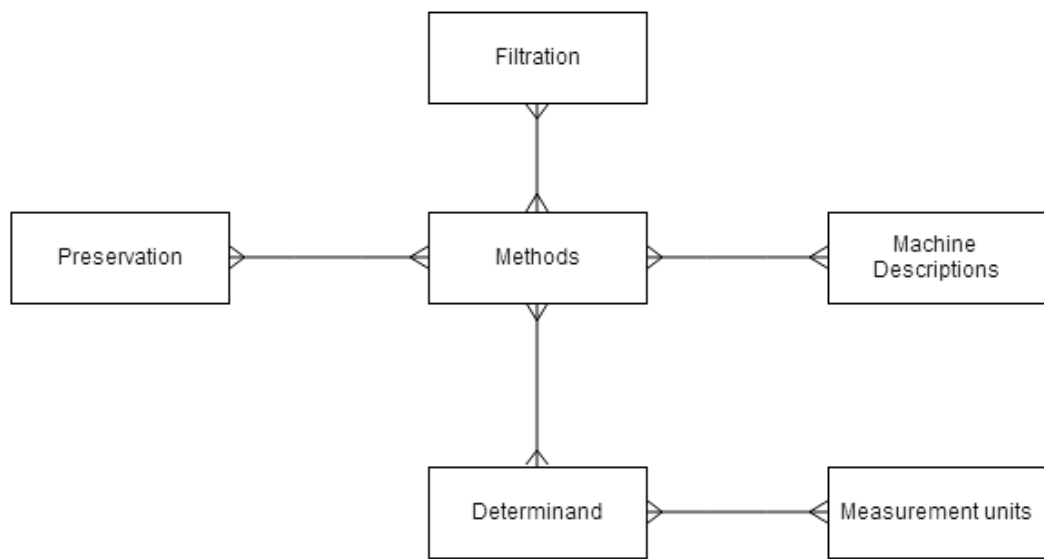




Figure 2.

onto.nerc.ac.uk/CAST/64.html



**Centre for Ecology & Hydrology**  
NATURAL ENVIRONMENT RESEARCH COUNCIL



# Vocabulary Service


HOME ADMINISTRATION

YOU ARE HERE: HOME > CAST > ACID RECOVERABLE POTASSIUM (EN)

CAST

Search the project

Overview Visual Browser SPARQL Endpoint

**acid recoverable potassium (en)** 

**Alternative Labels** *dissolved plus acid recoverable potassium (en)* *total K (en)* *total acid recoverable potassium (en)* *total acid available potassium (en)*

**Concept URI**  
<http://onto.nerc.ac.uk/CAST/64>

---

**Definitions**  
Total acid recoverable potassium, includes dissolved fraction plus that which was present in particulates but dissolved by acidification of the unfiltered sample to 1% with concentrated high purity nitric acid, without heating. This acidification takes place upon sampling and prior to cold storage and subsequent analysis. (en)

**Broader Concept**  
[determinands \(en\)](#)

**Narrower Concept**  
[dissolved potassium \(en\)](#)

NERC - Centre for Ecology & Hydrology