Characterisation of metadata to enable high quality climate applications and services

Deliverable 400.1

Data Model for “Commentary Metadata”

CHARMe is funded by the EC under its FP7 Research Programme

Document Control

Contributors

Person	Role	Organisation	Contribution
P J Kershaw	WP leader	STFC	Main author
S Ventouras	Data modelling	STFC	Consultation on modelling
M Nagni	Software development	STFC	Author provenance
A Wilson	Software development	STFC	Versioning provenance for annotations
S Blower	WP700 leader	University of Reading	Modelling for citation and fine-grained commentary
R Alegre	Software development	University of Reading	Modelling for fine-grained commentary
F Kratzenstein	Data modelling	DWD	Annotation target data types
I Rozum	Data modelling	ECMWF	Schema for Significant Events

Document Approval

Person	Role	Organisation

List of Acronyms

CEDA	Centre for Environmental Data Archival, RAL Space, STFC Rutherford Appleton Laboratory
CHARMe	CHARacterisation of Metadata
CiTO	Citation Typing Ontology, part of the SPAR (Semantic Publishing and Referencing) group of ontologies
CMIP5	Coupled Model Intercomparison Project Phase 5
CrossRef	Organisation with oversight for creation and maintenance of DOIs for publications
DataCite	Organisation with oversight for creation and maintenance of DOIs for datasets
DOI	Digital Object Identifier
ERS-1	ESA remote sensing satellite
ESGF	Earth System Grid Federation
FaBiO	Bibliographic ontology, a part of the SPAR (Semantic Publishing and Referencing) group of ontologies
FOAF	Friend Of A Friend ontology
GCMD	Global Change Master Directory – NASA Earth science metadata inventory

GEMET	GEneral Multilingual Environmental Thesaurus – a multilingual thesaurus that aims to define a core terminology for the environmental domain.
JSON	Javascript Simple Object Notation
MOLES	Metadata Object for Linking Environmental Sciences – Data Model for expressing 'B' metadata
OA	Open Annotation
REST	REpresentational State Transfer
SAR	Synthetic Aperture Radar
SRD	Software Requirements Document
TBD	To Be Defined
UC	Use Case
URD	User Requirements Document
URI	Uniform Resource Identifier
XML	eXtensible Mark-up Language

References

ID	Author	Document Title	Date
[R-1]	Project partners	Characterisation of metadata to enable high-quality climate applications and services	23rd November 2011
[R-2]	B.N Lawrence, R Lowry, P Miller, H Snaith and A Woolf.	Information in environmental data grids, Phil. Trans. R. Soc. A vol. 367 no. 1890 1003-1014	13th March 2009
[R-3]	STFC	Analysis of Existing Technical Solutions	14th May 2013
[R-4]	University of Reading	User Requirements, version 2.0	
[R-5]	CGI	System Requirements	TBD
[R-6]	CSIRO	Draft Ontology for ISO19115:2003, http://def.seegrid.csiro.au/isotc211/iso19115/2003/metadata/index.html	
[R-7]	CSIRO	Draft Ontology for ISO19156 (Observation and Sampling): http://def.seegrid.csiro.au/isotc211/iso19156/2011/	
[R-8]	Robert Sanderson, Paolo Ciccarese, Herbert Van	Open Annotation Data Model, Community Draft, http://www.openannotation.org/spec/core/	8th February 2013

ID	Author	Document Title	Date
	de Sompel		
[R-9]	Project partners	Characterisation of metadata to enable high-quality climate applications, Annex 1 Description of Work	2nd October 2012
[R-10]	Matthew Perry and John Herring	OGC GeoSPARQL - A Geographic Query Language for RDF Data, version 1.0	10 September 2012
[R-11]	Timothy Lebo, Satya Sahoo, Deborah McGuinness	PROV Ontology (http://www.w3.org/TR/prov-o/)	30 April 2013

Revision History

Issue	Author	Date	Description
0.1	P J Kershaw	2 September 2013	Internal Draft.
0.2	P J Kershaw	29 September 2013	Initial release to project partners Includes input from STFC team - Spiros Ventouras and Maurizio Nagni
0.3	P J Kershaw	14 November 2013	Fixed section numbering. Addressed initial feedback from partners: • Revised introduction • Increased size of diagrams and text in figures • Fixed citation example to label target as a dataset not metadata document • Simplified structure for tagging example
1.0	P J Kershaw	18 November 2013	First formal release. Updated following last round of review comments from project partners. • Extended introduction to provide more information about scope. • Explicit mention of WPs that are in scope for guidelines section 6. • Expanded Faceted Search and tagging to include more

Issue	Author	Date	Description
			information on candidate facets
1.1 Internal draft 3	P J Kershaw	22 October 2014	Additions for WP700 advanced uses of CHARMe – added additional content for Fine-grained Commentary, Significant Events. • Filled out data types – definitions needed for WP500 integration work. • New proposal for modelling citations. • Updated modelling for Significant events and Fine-grained commentary based on input for Raquel Alegre, Jon Blower and Adam Leadbetter (BODC) • Updated Provenance section to describe behaviours and relationships for deleted and modified annotations • Added Significant Events example Turtle • Added Reference to new CHARMe ontologies IRI in section 4 and 5.3
1.1 Issue	P J Kershaw	2 December 2014	• Added reference to CHARMe model code repository
1.2 Issue	P J Kershaw	17 April 2015	• Added references to new vocabularies for CF calendar types and vertical regions in 5.3.1 as now defined in NERC Vocabulary Server. With thanks to Roy Lowry BODC and Alison Pamment STFC.

Table of Contents

1 Introduction

This document sets out the conceptual data model for Commentary (C) metadata for CHARMe. The target domain for the model is Earth observation and climate research. However, there are common elements that could equally be applied to a broader range of scientific or research disciplines. In the description of work [R-9] it states, "The data model needs to take into account the diversity of the data sets ranging from satellite data (raw data as well as satellite derived climate data records), in-situ data to model data (e.g. reanalysis data)." This a central challenge to CHARMe: the variety and complexity of data in scope make it difficult to represent every possible use case. There is a danger that this is translated into the model making it too complicated or prescriptive. The alternative is to make a something more generic which is flexible enough to support a broad scope supported through specialisations to meet the needs of individual use cases. This is the approach followed here.

The document structure is organised as a progression of steps tracing through from the requirements gathering and analysis through to it application in specific cases and guidelines for integration with existing data providers:

1. A review of the definition of C-metadata and its scope drawing from the prior work in the proposal document [R-1] and Analysis of Existing Technical Solutions [R-3]
2. A presentation of Open Annotation, the proposed basis for CHARMe's data model.
3. Analysis of the information types in the system drawing from the requirements gathering phase
4. Application of the Open Annotation model to address the use cases and the information types identified for CHARMe's domain of interest.
5. Profiling and extension of the model to meet specific applications identified in WP700, *Application to Climate Services*.
6. Identify principles for integrating C-metadata with existing data holdings at data providers

Points 4) and 5) relate to the scope of the model. Item 4) covered in section 5.2, covers the simplest use cases such as for example making a text comment about a target. 5) is explored in section 5.3. These are to varying extents specialisations of the basic model. For example, the intercomparison of datasets is likely to involve a set of target properties applicable to only a very specialised subset of data types. Faceted search it could be argued is generically useful across a range of use cases. Nevertheless, defining a suitable set of facets across a broad domain of usage may be difficult and require customisation to meet specific needs. At the time of writing these aspects of the model are open to further development as WP700 progresses whereas the core set out in section 5.2 should be considered as fixed.

A distinction should be drawn between this document and the related WP430 deliverable. This document is concerned with the model alone, whereas the latter defines how the model is translated into machine-readable form. Given the Linked Data approach adopted, the various aspects of the model are depicted using RDF graphs and Turtle format. Turtle is used because it provides a simple human readable expression for RDF-triples. It has no relation to the encodings to be used.

The reference URI for the data model for CHARM is http://purl.org/voc/charme. The classes and vocabularies for the model are maintained at https://github.com/cedadev/charme-data-model/.

2 Review of C-Metadata Definition

The proposal document [R-1] specified the scope of Commentary (C) metadata within the domain of climate science and Earth observation:

1. Post-fact annotations, e.g. citations, ad-hoc comments and notes;
2. Results of assessments, e.g. validation campaigns, intercomparisons with models or other observations, reanalysis;
3. Provenance, e.g. dependencies on other datasets, processing algorithms and chain, data source;
4. Properties of data distribution, e.g. data policy and licensing, timeliness (is the data delivered in real time?), reliability;
5. External events that may affect the data, e.g. volcanic eruptions, El-Nino index, satellite or instrument failure, operational changes to the orbit calculations.

3) and 4) could be more correctly labeled as B-level metadata[1] as defined by Lawrence et. al. [R-2]. These items were included following user feedback in the requirements gathering phase. This is a reflection of users' experience that many data providers fail to provide these clearly in their metadata. However, it also highlights a model for how a CHARMe system could operate as a virtuous circle. - Users annotate existing data provider metadata with additional information to augment it. Additional information provided in these annotations can then be fed back by the data providers to improve their metadata catalogues.

Important discriminating factors here are the author of the metadata - e.g. a data provider, a data centre or some third party. The former two have some form of authority over the data whereas the latter does not. A second factor is the time of creation of the metadata, for example, at the time of curation at a data centre or some later time.

3 Review of the Open Annotation Data Model

This section is intended to provide a brief review of the OA model [R-8]. This is followed up by an examination of how to best apply it to CHARMe. An important conclusion from the analysis of existing technical solutions [R-3] was to adopt a Linked Data approach for CHARMe and use OA as an overall framework.

OA is concerned with the *annotation* of data with additional information. Core to the model are these basic concepts:

- **target** – the subject of an annotation
- **body** – a comment, classification or another resource which is to be associated with the target
- **annotation** - the conceptual linkage between the two

[1] CHARMe mailing list discussion 5th September 2013

[2] Ref: http://www.w3.org/DesignIssues/LinkedData.html

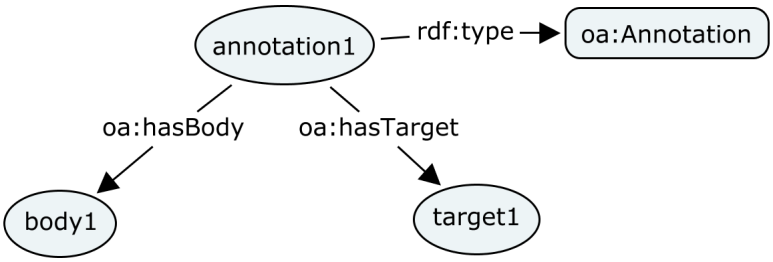

Figure 1: Open Annotation - basic concepts

Body and target can be typed but the model deliberately avoids defining explicit types for these. This allows flexibility since for example the body of one annotation may be the target of another. This enables the building of a chain of linked annotations much in the same way as a discussion thread used on mailing lists or web forums.

A number of other variations are possible to meet different requirements: a given annotation can have more than one target, more than one body or even no body at all. Target and bodies may be as simple as links to existing web-based resources such as documents, images or datasets for example, or can be new objects created for the annotation. The obvious example is an annotation that is a comment about a given target. Here the annotation body contains the comment text. More complex associations may be made via *Specifiers*, a means of describing a subset of a given resource.

The `oa:Annotation` class enables the association of metadata including provenance data and a controlled vocabulary to identify the motivation for creating the annotation. This is extended in the case of *tagging*. This enables annotations to be classified, associating them with a given text keyword or term from a vocabulary specified in the annotation body.

4 Analysis of Information Types

An analysis of the types of information to be supported provides a basis for the model. This has been drawn from the following sources:

1. The proposal document: initial analysis to address the project call [R-1]
2. D300.2 Analysis of Existing Technical Solutions [R-3]
3. The URD [R-4]
4. Results from discussions with the project partners and requests for feedback

It is worth expanding 4) to clarify: a request was made for partners who are data providers to list the types of grey literature that they wish to include in the system. Gathering these various sources together, the following table lists these (left hand column) against a number of key attributes (subsequent columns). The latter are significant factors in determining how the information is modelled. The contents of these columns are filled as follows:

Yes = bold statement in column heading is true
No = bold statement in column heading is false
? = unclear which is true, else true in some cases, false in others

Information Type	**Internal** or external to system	**Structured** or Unstructured	**Existing** or proposed	Notes
Datasets	No	Yes	Yes	The exact definition of what constitutes a dataset can be arbitrary but can be generally agreed upon within a specific community or domain.
Dataset Collections	No	No	Yes	According to the type of data, the domain, different hierarchies may be applied. For example, ISO19115 has the concept of a Dataset Series. CEDA MOLES defines an ObservationCollection.
Discovery Metadata	No	Yes	Yes	ISO19115 and GCMD DIF are examples
Browse Metadata	No	Yes	Yes	CEDA MOLES
Instrument information	No	No	Yes	Dataset collections may provide an alternative identifier for an instrument e.g. collection of ERS-1 SAR data
Algorithm Theoretical Basis Documents (ATBDs)	No	No	Yes	
Product User Manuals (PUMs)	No	No	Yes	
Validation Reports	No	?	Yes	
Operations Reports	No	?	?	Specific to a data provider or instrument operator
Service Messages	No	?	?	Specific to a data provider
Product Change Logs	?	?	?	Specific to a data provider
Known Product Disruptions	?	?	?	A specialisation of a significant event
Significant Events	?	Yes	No	Required for work package 720
Ad hoc comments and notes	Yes	No	No	These annotations themselves can be the subject of other annotations
Papers	No	No	Yes	
DOI metadata	No	Yes	Yes	DataCite and CrossRef provide a means to access this by resolving the DOI with an XML or JSON Accept Header to return the respective content.

Table 1: Analysis of Information Types

Reviewing the content:

- Nearly all the types identified are external to the CHARMe system and already in existence.
- The majority of types are unstructured
- Significant Events are notable in that they are new, likely to be structured and *could* be stored internally within a CHARMe node.

Building on Linked Data principles[2], the first requirement for these information types is that instances MUST be HTTP URIs, enabling them to be uniquely identified and resolved. For structured content, a second question arises regarding their serialisation. Representation as RDF enables further possibilities for linking and discovery within CHARMe and other similar systems. However, the potential benefits needed to be weighed carefully against the implementation effort especially with large legacy systems.

These examples could equally form bodies or targets for annotations. The following table classifies them into proposed types. This makes use of FaBiO[3] from the SPAR family of ontologies. FaBiO provides a relevant list of types covering reports, papers and metadata documents. There is scope to provide more fine-grained definition but a balance needs to be struck. If the terms are too specific they may be difficult to discover or match with related items. Where types are omitted, these are either specific to a given data provider or for which further analysis will be required in other work packages.

Item	Proposed type	Notes
Datasets	dctypes:Dataset	May be possible to refine for specific use cases
Dataset Collections	fabio:MetadataDocument	Draft ontologies for ISO19115 and ISO19156 could be applied here [R-6][R-7] e.g. ISO 19115 DatasetSeries. Collections can also be represented natively with Open Annotation using multiple targets to a body or oa:Composite class.[4]
Discovery Metadata		
Browse Metadata		
Instrument information	?	Not defined currently but Airbus have a use case for this.
Algorithm Theoretical Basis Documents (ATBDs)	charme:AlgorithmTheoreticalBasisDocument	
Product User Manuals (PUMs)	charme:ProductUserManual	
Validation Reports	charme:ValidationReport	
Operations Reports	charme:OperationReport	
Service Messages	charme:ServiceMessage	Depend on specific implementations
Product Change Logs	charme:ProductChangeLog	
Known Product Disruptions	charme:KnownProductDisruption	
Significant Events	charme:SignificantEvent	New type, for WP720 requirement

[2] http://www.w3.org/DesignIssues/LinkedData.html
[3] http://purl.org/spar/fabio
[4] http://openannotation.org/spec/core/multiplicity.html#Composite

Ad hoc comments and notes	cnt:ContentAsText, cnt:ContentAsXML	As recommended by OA, cnt:ContentAsXML suitable for XHTML.
Papers	fabio:ResearchPaper, fabio:ConferencePaper	
DOI metadata	rdf:description	Following form provided by CrossRef and DataCite

Table 2: Proposed Typing for Information Types

Referenced qualified names are:

- dctypes: http://purl.org/dc/dcmitype/
- fabio: http://purl.org/spar/fabio/
- rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
- cnt: http://www.w3.org/TR/Content-in-RDF10/
- charme: http://purl.org/voc/charme

A SKOS vocabulary has been created to define the new CHARMe-specific types.

5 Application of Open Annotation Data Model to CHARMe

Open Annotation provides the framework for expressing relationships for the information types identified in the previous section. Depending on the assertion being made, instances of each of these types may be expressed as annotation bodies or targets. This section contains a number of examples, first the core use cases and then in a separate sub-section, specific application for the four areas identified in WP700. At the time of writing, WP700 is still in the requirements phase so the content here is intended as an initial analysis that will be subject to development and revision as this work progresses.

Distinct CHARMe-specific types are required as specified in "Table 2: Proposed Typing for Information Types" above. The elements required can be represented from OA and existing ontologies. The table below lists the ontologies used in the modelling.

Name	Release	URI
Open Annotation	0.9.20130208	http://www.w3.org/ns/oa
RDF	1999/02/22	http://www.w3.org/1999/02/22-rdf-syntax-ns#
OWL	2002/07	http://www.w3.org/2002/07/owl#
Dublin Core Metadata Initiative types	2012-06-14	http://purl.org/dc/dcmitype/
Dublin Core Metadata Initiative elements	1.1	http://purl.org/dc/elements/1.1/
FOAF	0.1	http://xmlns.com/foaf/0.1/
Representing Content in RDF	1.0	http://www.w3.org/TR/Content-in-RDF10/
CiTO SPAR Ontology	2.6.2	http://purl.org/spar/cito/
FaBiO SPAR Ontology	1.7.5	http://purl.org/spar/fabio/

Table 3: Ontologies Applied in Modelling

5.1 Annotation Core Properties

These are key properties that are common to all annotations. They are associated with the `oa:Annotation` class.

5.1.1 Motivations

Motivations provide important context information answering the question why a given annotation was created. The model provides a class `oa:Motivation` derived from `skos:Concept`. A controlled vocabulary defines the possible motivations. These are applicable in the context of CHARMe. However, there may be cases where more specialised motivations are needed. Rather than extend the existing class, the OA specification ([R-8], Appendix B) recommends adding new motivations by *relating* them to existing ones using SKOS predicates such as `skos:broader` or `skos:narrower`. One example could be citation. OA does not provide a specific motivation for citation. The closest match is oa:linking. However, this could be refined to a more specific citing motivation either defined as a new motivation or using terms from the CiTO ontology. The citation use case is examined more detail in section 5.2.2.

`oa:tagging` is worth mentioning explicitly as this has flags an annotation as having a special body type used for tagging, an `oa:Tag` or `oa:SemanticTag` type. This is discussed further in the following section, 5.3.5.

5.1.2 Provenance

Provenance includes the authorship of annotations and the creating agent, the associated date/times for creation and also, versioning information.

5.1.3 Authorship

Considering authorship first, the OA model allows for an `oa:annotatedBy` property of an annotation.[5] OA stipulates that there should be at least one `oa:annotatedBy` but there may be none or more than one. The use of the FOAF ontology[6] is recommended.

For CHARMe, there MUST be one or more authors and this is expressed with FOAF. This is expected to be individual users in which case the `foaf:Person` class SHOULD be used. However, there MAY be cases where `foaf:Organization` is more appropriate, for example, for bulk creation of annotations by a program at a data provider site.

- `foaf:name` is used as the descriptive, human readable identifier. This is not guaranteed to be unique.
- `foaf:mbox` - an e-mail address MAY be provided as a contact for the user. For reasons of user privacy this is not mandatory.

The URI for the Person object provides a unique identifier.

```
@foaf: <http://xmlns.com/foaf/0.1/> .


<person1> a foaf:Person ;
       foaf:givenName “Maurizio” ;
       foaf:familyName “Nagni” ;
       foaf:account <account1> ;
```

[5] http://www.openannotation.org/spec/core/core.html#Provenance
[6] http://xmlns.com/foaf/spec/


```
        foaf:mbox M.Nagni@somewhere.ac.uk .
```

Figure 2: FOAF attributes for oa:annotatedBy

The nature and requirements for author information have been arrived at after careful consideration. The system needs to be able to attribute authorship and notify users of changes to annotations[7]. The model should also strike a balance between measures to prevent misuse of the system but it should also protect user privacy. Handling of e-mail addresses is key:

- Users provide their e-mail address to the CHARMe node *on registration* with the CHARMe node. They do so *optionally* in order that they can be notified of changes to annotations they have authored. This information is **private** and only for the purpose of allowing the node to notify the author of changes.
- When a given author creates a new annotation, they can *optionally* provide an e-mail address for inclusion in the `oa:annotatedBy foaf:Person` object. This information is **public**. They provide this to give other users a contact point associated with that annotation. This could be for example an e-mail address for a helpdesk where others users can make further enquiries.

When a new annotation is submitted, the node sets a unique URI for the `foaf:Person` object submitted by the client. This acts as a primary key between the node's internal user records and the annotations submitted enabling a direct link between annotation author and their e-mail address.

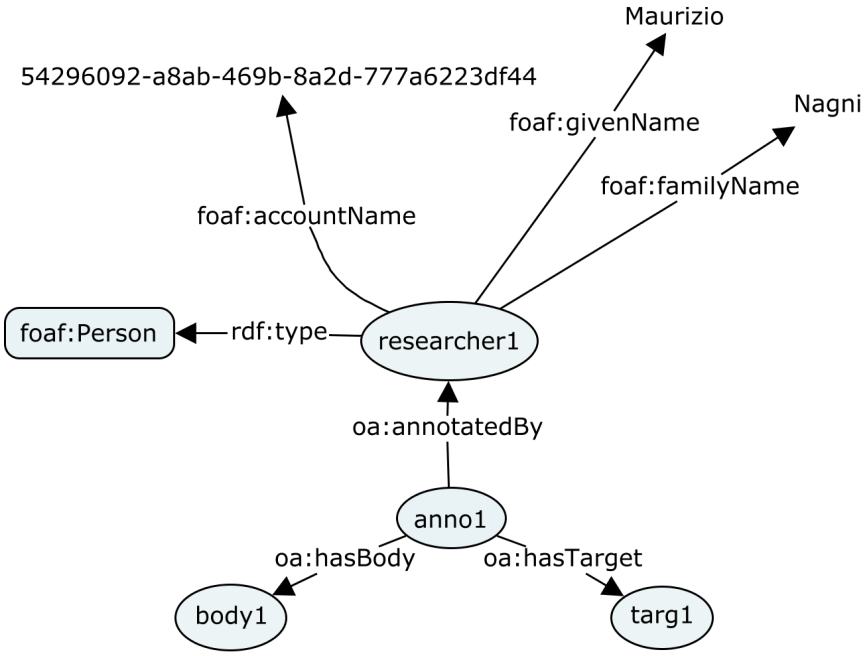

Figure 3: Authorship of Annotation

5.1.4 Creating Agent

The second aspect of provenance to consider, the creating agent can be captured with the `oa:serializedBy` property. This can be used to set metadata for the software agent

[7] Requirement number R-116, URD [R-4]

creating or *updating* a given annotation. This MAY provide useful contextual information. Further discussion is needed amongst the project partners to agree its usage.

5.1.5 Management of Modification and Deletion of Annotations

The ability to modify or delete existing annotations is an important requirement particularly, for example if a user submits an annotation and realises that they have made a mistake and wish to correct it or if a moderator wishes to delete inappropriate material.

CHARMe organises data in a node into two main graphs named *submitted* and *retired*. New annotations are added to the submitted graph. When a request is made to delete an annotation it is re-assigned from the submitted to the retired graph. The annotation is effectively retained in the triple store but is moved to a separate graph to indicate its new status.

Modification of an annotation presents some challenges semantically. Should an existing annotation be modified, it potentially now has a completely new meaning and context. This has implications for other annotations or nodes that reference it. Imagine an annotation A created by user Alice asserting that a given dataset has an incorrect calibration. User Ben comments on this annotation and refutes this with a comment in annotation B. Alice then realises that she has made a mistake and modifying annotation A, changes the body text to correct it. However, Ben's annotation still remains and makes no sense now since it was based on Alice's original assertion, not on the corrected version of the annotation.

In order to enable traceability, a modification operation creates a new annotation in the submitted graph and moves the annotation it replaces to the retired graph. This addresses the case above where an annotation is modified which has already been deleted. Ben's annotation still points to the deleted annotation that he originally commented on. However, it now has a deleted status so it is clear that it has been superseded by a new version.

This still leaves the problem of how to trace between a modified annotation and the one it replaced. Open Annotation has a provenance model that applies a simplified approach to the PROV-O [R-11] model from which it derives. It is possible to set an `oa:annotatedBy` predicate to an annotation but if an annotation is a modified version of an original, there is no way to represent this with Open Annotation. Since Open Annotation uses PROV-O, though the latter can be readily exploited to represent provenance information associated with a change or modification:

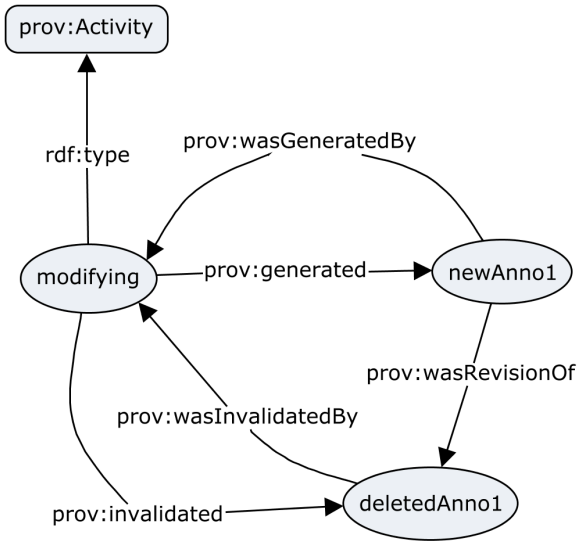

Figure 4: Provenance for Modification of an Annotation

The new annotation can be identified as a modification since it has a `prov:wasRevisionOf` predicate pointing to the deleted annotation it replaces. The deleted annotation likewise can be identified as a deletion with the `prov:wasInvalidatedBy` predicate. This predicate points to the *Activity* of modifying the annotation rather than directly back to `newAnno1`. In this way it is possible to differentiate between a deleted annotation and one that has been deleted as the result of a modification. In the case of the latter, the modifying activity points back to the new annotation via `prov:generated`. In addition, author information – via FOAF - and creation time can be added to the modifying activity – using `prov:startedAt` and `prov:endedAt` predicates.

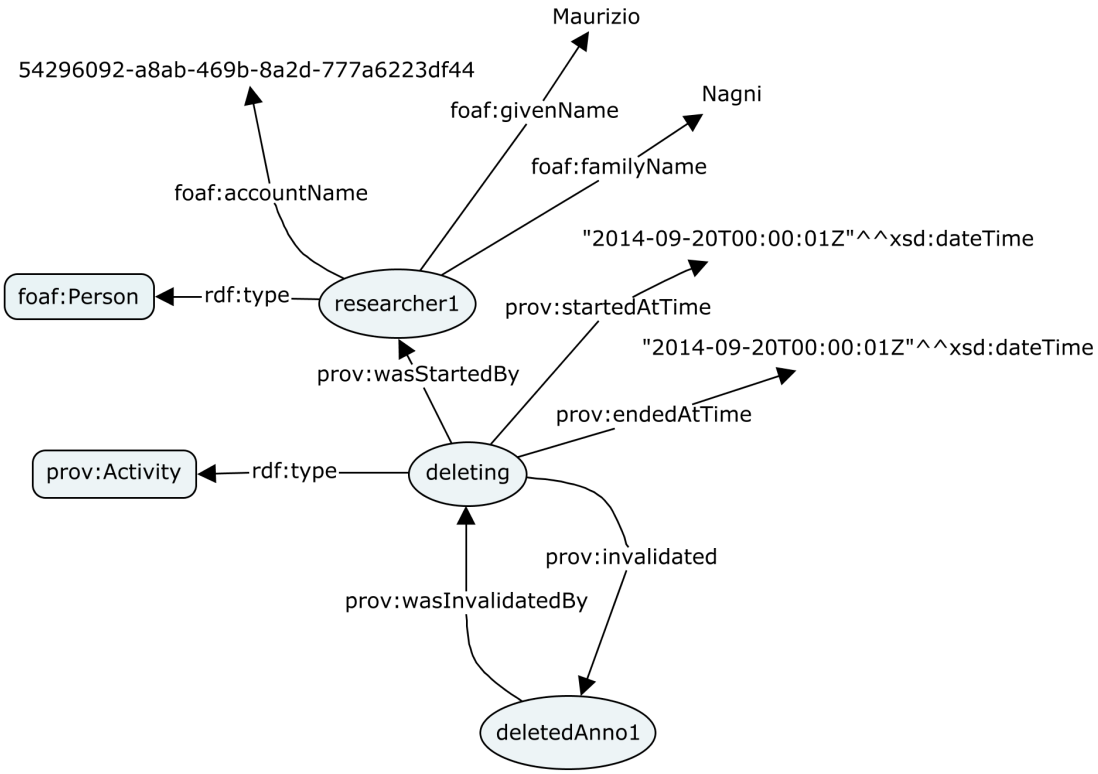

Figure 5: Provenance Information for deleted annotation

5.2 Core Annotation Types

5.2.1 Text-based Annotations

This is the simple case where a user writes a free text comment about a given target. The text is created and attached as a new annotation body. Unstructured plain text is stored as `cnt:ContentAsText`[8]. Structured content is also possible via `cnt:ContentAsXML` for representing for example XHTML.

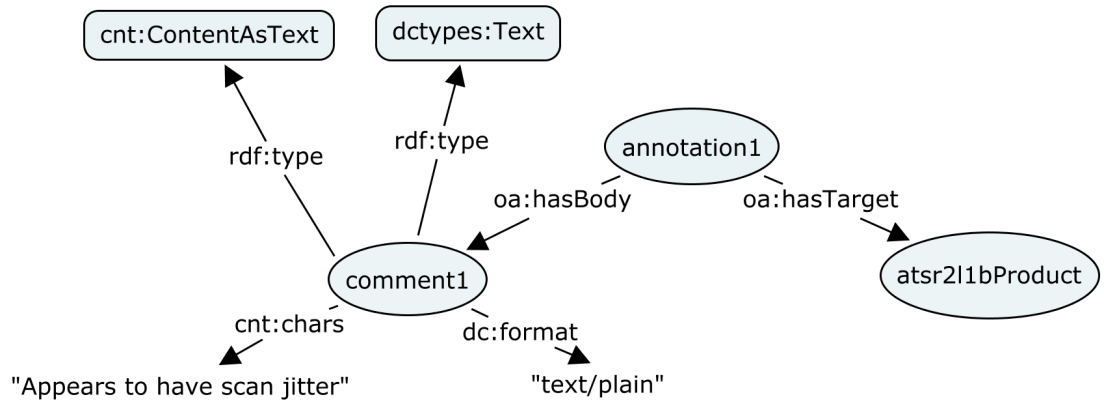

Figure 6: Text Comment

5.2.2 Citation

A natural extension of the concept of a document linking to a dataset is a paper citing a dataset. Here we can use the CiTO ontology together with OA to express this relationship in more concrete terms:

```
@prefix oa: <http://www.openannotation.org/spec/core/> .
@prefix dctypes: < http://purl.org/dc/dcmitype/> .
@prefix cito: <http://purl.org/spar/cito/> .

<citation1> a oa:Annotation ;
       a cito:CitationAct ;
       oa:hasBody <citation> ;
       oa:hasTarget <http://dx.doi.org/10.5285/aef749ed-3ee9-4ce0-a0f9-
       e5bb369b5861> ;
       oa:motivatedBy oa:linking .
       cito:hasCitingEntity <http://dx.doi.org/10.5285/4BCFA3A4-C7EC-4414-
       863D-CAECEB21F16F> ;
       cito:hasCitationEvent cito:citesAsDataSource ;
       cito:hasCitedEntity < http://dx.doi.org/10.5285/aef749ed-3ee9-4ce0-
       a0f9-e5bb369b5861> .

<http://dx.doi.org/10.5285/4BCFA3A4-C7EC-4414-863D-CAECEB21F16F> a
fabio:ConferencePaper .
< http://dx.doi.org/10.5285/aef749ed-3ee9-4ce0-a0f9-e5bb369b5861> a
dctypes:Dataset .
```

Figure 7: Turtle – Linking a dataset to publication that cites it

[8] http://www.w3.org/TR/Content-in-RDF10/

Note that the entry from the CEDA MOLES catalogue makes use of the FaBiO ontology. It could be labelled as an *Observation* in CEDA MOLES 3 terminology. This is based on O&M, so it should be possible to reference terms direct from CSIRO's draft O&M ontology [R-6].

CrossRef and DataCite provide standardised metadata for DOIs. This snippet is extracted from the DOI metadata for the citing paper mentioned above. This information can be linked via the `sameAs` relationship matching with the paper's DOI included in the `CitationAct cito:hasCitingEntity` property. Although the information is available directly from the DOI, we propose caching this in the CHARMe repository to avoid the need to explicitly pull the information from an external source on subsequent requests.

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<paperDescr1> a rdf:Description ;
    dc:creator “Haines, Keith” ;
    dc:creator “Valdivisio, Maria” ;
    dc:publisher “NERC British Atmospheric Data Centre” ;
    dc:title “Global Ocean Physics Reanalysis UR025.4 (1989-2010) as part of
the VALue of the RAPID-WATCH Climate Change programme array (VALOR) project”
;
    dc:date “2013” ;
    owl:sameAs <http://dx.doi.org/10.5285/4BCFA3A4-C7EC-4414-863D-
CAECEB21F16F> ;
    dc:identifier “10.5285/4BCFA3A4-C7EC-4414-863D-CAECEB21F16F” .
```

Figure 8: DOI Metadata

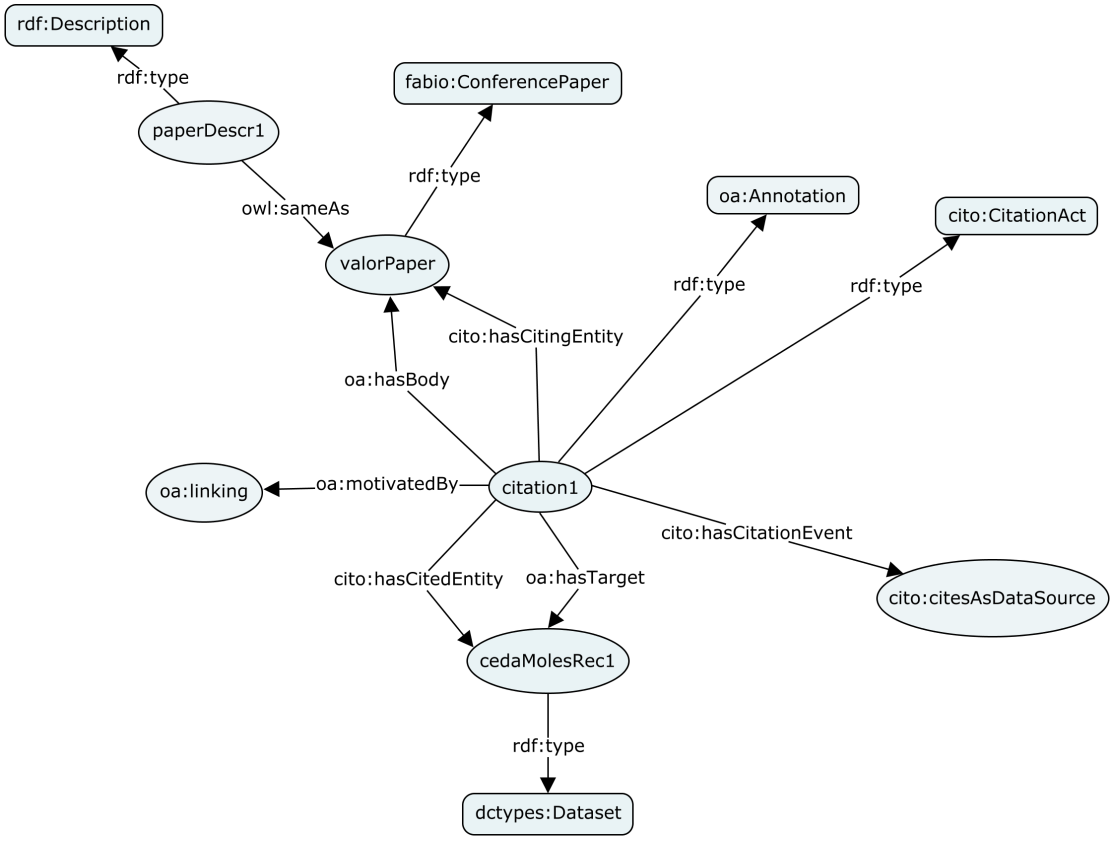

Figure 9: Linking a dataset to publication that cites it

The graph in the above figure illustrates the relationship between features from the OA and CiTO models. This presents a change from the original model proposed in version 1.0 of the data model. Rather than represent independent citation and annotation objects, the citation1 object above is both an `oa:Annotation` and a `cito:CitationAct`. It has predicates for both. In this way it is compatible with reasoners that are aware of either of the two models.

5.3 Profiling and Extensions

This section examines the application of the CHARMe model to the specific application areas identified in work package 700. This will require extension of the model to include new information types and profiling to define the particular scope or domain of use. A new charme IRI has been defined for specific ontologies created for CHARMe:

http://purl.org/voc/charme

Work package 700 provides a starting point from which to evaluate the model against practical use cases. These are not exhaustive and further use cases are expected outside the scope of the current project as the system is deployed and developed into the future.

5.3.1 Annotation of a Target Subset: Fine-grained Commentary

OA's Selectors provide the basis for defining subsets of resources as required for fine-grained Commentary from WP740. The model provides specific examples but these are focused on text and image-based subsets rather than more specialist geo-spatial data. The

DataCube ontology has been proposed [R-3] as one solution for the representation of such subsets. However, this is more suited to discrete locations rather arbitrary coverages.

The diagram below shows the current proposal. This builds on Open Annotation's Specifiers and Specific Resources module[9]. This enables a subset of a given target to be specified. A number of selectors are defined including an SVG one to select an area. However, there are no types for specifying geo-temporal bounds. For CHARMe then, we create a new SpecificResource, charme:DatasetSubset derived from oa:SpecificResource. This acts as a proxy to the dataset subset defining the source dataset and the subset specified. A selector `charme:SubsetSelector`. The subset includes temporal bounds and includes the ability to associate it with a specific calendar. This is relevant to some model data that uses non-specific calendars. Working with the BODC (British Oceanographic Data Centre), calendars have been defined as a new SKOS vocabulary based on the names defined in CF and incorporated in the NERC Vocabulary Server - "Climate and Forecast Calendars" - http://vocab.ndg.nerc.ac.uk/list/P370/current.

Geospatial extent is set with three distinct categories: vertical extent, a horizontal extent and a named region. Separating horizontal and vertical extent allows the re-use of existing predicates and classes from the GIS domain where vertical extent is often not dealt with explicitly. Named region refers to standard vocabularies of named regions on the globe such as for example, '`North Atlantic`'. In the example below, geographic extent is expressed as both a named region and a polygon. In practice the two would likely to be exclusive of one another. The polygon is defined with predicates from GeoSPARQL [R-10]. The geoid defaults to WGS84. Vertical extents can be expressed as a range or a named region such as stratosphere. For the former, it unlikely that this will be defined within the scope of the current project. Nevertheless, a placeholder is there in order to incorporate future enhancements. Vertical named regions can be named uses region names defined in CF names. These have been as a new SKOS vocabulary served from the NERC Vocabulary server as, "Climate and Forecast Vertical Co-ordinate Coverages" - http://vocab.ndg.nerc.ac.uk/list/P380/current. This work is again thanks to the BODC for their collaboration with this effort.

As well as geo-temporal bounds it is also possible to define along the axis of variable name using the `charme:hasVariable` predicate. This could points to a specific `charme:Variable` class which can define before a internal variable name specific to a dataset file and the associated CF standard name if appropriate. Both properties are required since it must cater for the case where there is more than one variable in a dataset with the same standard name.

[9] http://openannotation.org/spec/core/specific.html

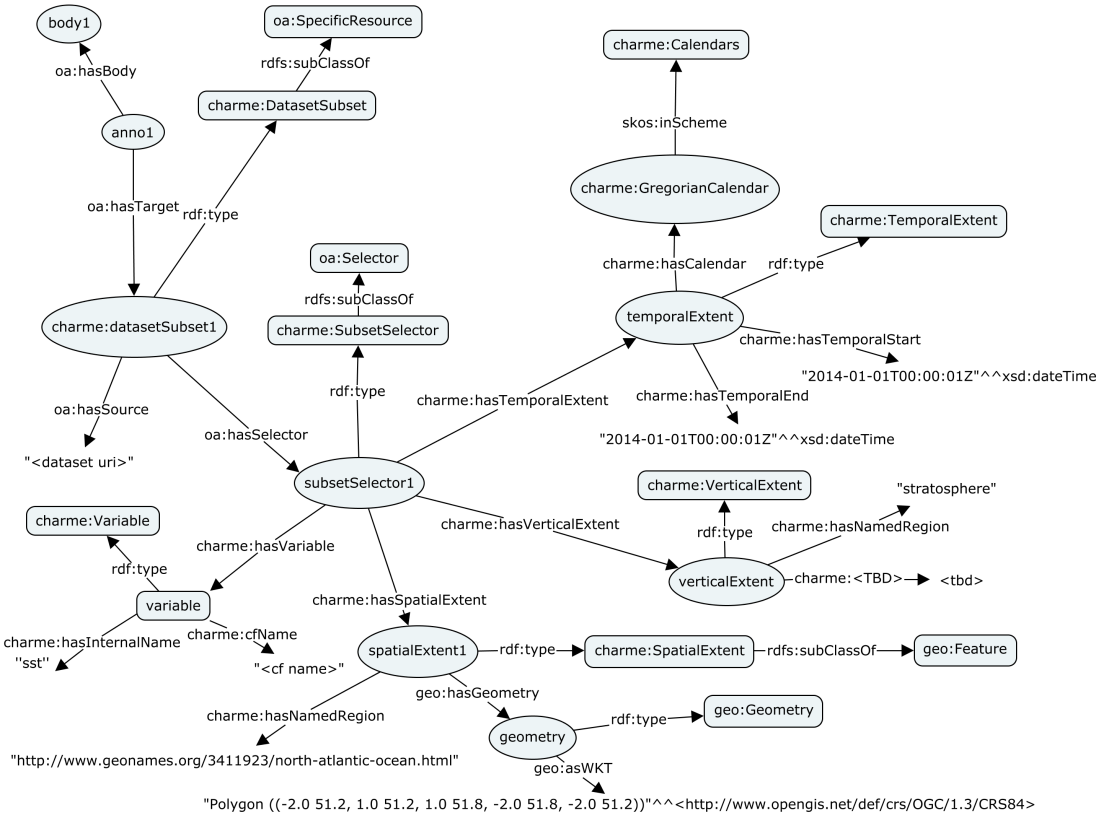

Figure 10: Fine-Grained Commentary

5.3.2 Intercomparison of Datasets

This task, WP730 will require resources with specific structured metadata according to the types of parameter to be inter-compared. This is largely out of scope of the core CHARMe metadata and will require profiling of the model as part of this task.

5.3.3 Significant Events

One of the tasks of work package 700 is the development of a *significant events* viewer. The specification of the Significant Event type will be part of this work. This section provides some initial analysis that may be subject to change and/or refinement.

The characteristics of a Significant Event are:

1. A location or geographic extent
2. A time or time period
3. A description of the event

Since these characteristics are concerned with geographic information it could be modelled as a *Feature Type* according to ISO19100 series and the more recent GeoSPARQL [R-10] specification. This enables us to take advantage of the established concepts in this series of standards and existing implementations.

ECMWF have modelled a Significant Event as holding these key pieces of information

Property	Type	Description
time_st	datetime	Time the event started
time_en	datetime	Time marking the end of the event
event_type	Code list item	Event type
event_subtype	Code list item	Sub-type for the event
event_name	char	Name of the event
event_summary	char	Event summary information
URL	char	URL to relevant information about the event
named_region	char	The geographic region for the event
country	char	Named country coincident with the event
latitude	double	Co-ordinate Reference System is WGS84
longitude	double	Co-ordinate Reference System is WGS84

Table 4: Mappings from Significant Events database schema

There are a number of existing ontologies that could be re-used or adapted for the representation of a Significant Event. Re-using existing ontologies enables consumption of the content by client that are already aware of these existing ontologies. Existing ontologies if already in use are more likely to have been tried and tested in practice and come under scrutiny to assess their suitability.

Two examples are:

- The Event Ontology
- PROV-O

PROV-O is suitable since OA builds on it as a means to extend provenance information about annotations and it is currently maintained. There are three main classes: Entity, Activity and Agent. Examining the role of these classes, the concept of a significant event maps closely to the Activity class.

Taking this further, we can express event type and event sub-types through two pathways: as concepts in a SKOS scheme or via class sub types using OWL[10] and RDFS[11]. For example, a hurricane event is a sub-type of climate event that is itself a sub-type of a significant event. In the same way, hurricane and climate events can be expressed as concepts in a Significant

[10] http://www.w3.org/TR/owl-ref/
[11] http://www.w3.org/TR/rdf-schema/

Events SKOS concept scheme. Climate event concept is a broader term than hurricane event.

Location information fields such as named region and country can be described with datatype properties derived from prov:atLocation. Latitude and longitude can be represented using GeoSPARQL [R-10]. `geo:Feature` and `geo:Geometry` classes allow the representation of multiple feature geometries in addition to single points. Finally string literals such as event name, summary and the information field can be derived from matching properties from other ontologies already used e.g. information maps rdfs:seeAlso.

The table below shows how the significant event database schema fields can map to PROV-O, SKOS and GeoSPARQL predicates and types.

Significant Event database table Field name	**Predicate**	**Type**	**Notes**
time_st	prov:startedAtTime	xsd:datetime	Time the event started
time_en	prov:endedAtTime	xsd:datetime	Time marking the end of the event
event_type	rdf:type	skos:Concept / owl:Class	Extend PROV-O Activity class and define SKOS concept in a significant events concept scheme
event_subtype	rdf:type	skos:Concept / owl:Class	Extend PROV-O Activity class and define SKOS concept in a significant events concept scheme
event_name	skos:prefLabel	String literal	Extend SKOS data type property
event_summary	skos:note	String literal	Set property with SignificantEvent class as rdfs:domain
information	rdfs:seeAlso	String literal	Set property with SignificantEvent class as rdfs:domain

named region	prov:atLocation	prov:Location / Geonames URI	The geographic region for the event. Use Geonames ontology
country	prov:atLocation	prov:Location / Geonames URI	Named country coincident with the event. Use Geonames ontology
latitude	prov:atLocation	prov:Location expressed as geo:Feature and sf:Point	sf:Point from http://www.opengis.net/ont/sf#
longitude	prov:atLocation	prov:Location expressed as geo:Feature and sf:Point	sf:Point from http://www.opengis.net/ont/sf#

Table 5: Mappings from Significant Events schema to ontologies

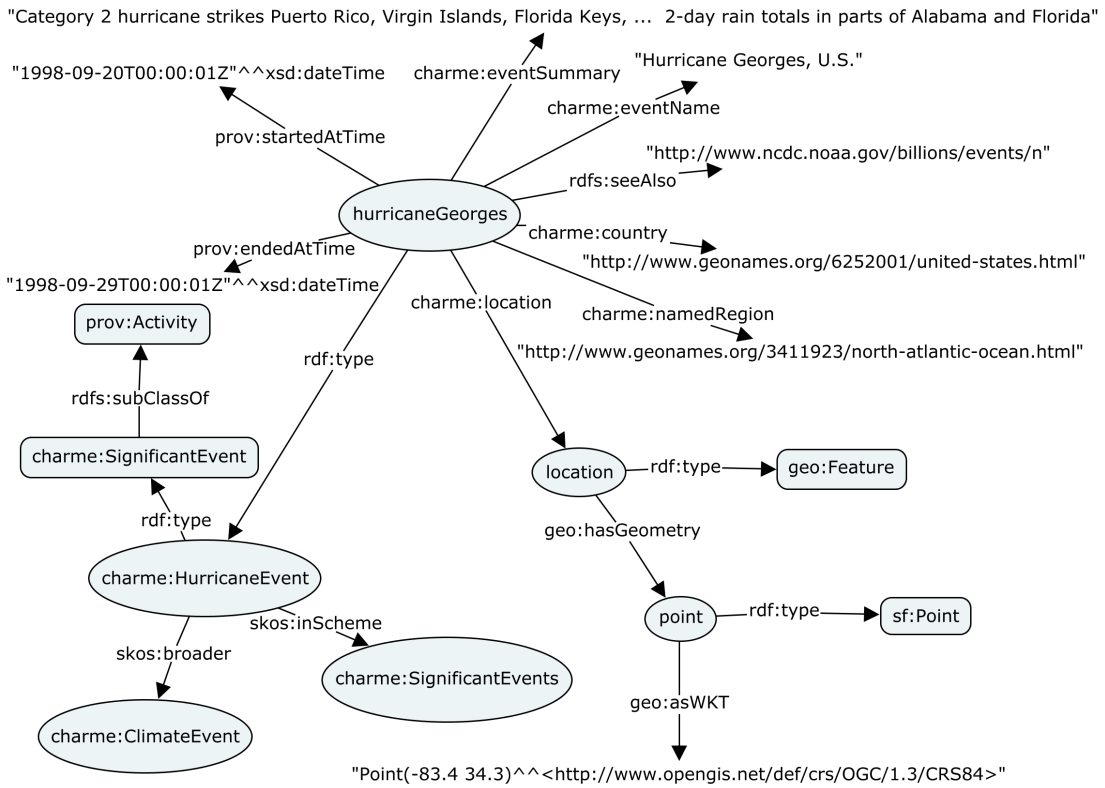

Figure 11: Significant Events Example Graph

```
@prefix charme: <http://purl.org/voc/charme#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
```

```
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix sf: <http://www.opengis.net/ont/sf#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<hurricaneGeorges> rdf:type charme:HurricaneEvent ;
    charme:eventName "Hurricane Georges, U.S." ;
    charme:eventSummary "Category 2 hurricane strikes Puerto Rico, Virgin
Islands, Florida Keys, ... 2-day rain totals in parts of Alabama and
Florida" ;
    rdfs:seeAlso "http://www.ncdc.noaa.gov/billions/events/n" ;
    charme:country "http://www.geonames.org/6252001/united-states.html" ;
    charme:namedRegion "http://www.geonames.org/3411923/north-atlantic-
ocean.html" ;
    charme:location <location1> ;
    prov:startedAtTime "1998-09-20T00:00:01Z"^^xsd:dateTime ;

5.3.4      prov:endedAtTime "1998-09-29T00:00:01Z"^^xsd:dateTime .

<location1> rdf:type geo:Feature ;
     geo:hasGeometry <point1> .

<point1> rdf:type sf:Point ;
     geo:asWKT "Point(-83.4
34.3)^^<http://www.opengis.net/def/crs/OGC/1.3/CRS84>" .
```

Figure : Significant Events Example Turtle

5.3.5 Faceted Search and Tagging

Faceted search enables users enhanced search capability enabling them to refine searches of the CHARMe repository by categories. In order to create a facet, the data must be organised by some corresponding discrete categorisation with a manageable number of values. Authorship of annotations is an example where there is a discrete category but the number of different values is not constrained and so is impracticable to implement as a facet.

The choice of facets to expose must be driven by user needs and not solely oriented by what the data structure dictates. These facets are proposed:

- Classification by domain keyword (see following discussion)
- Information type as specified in section 4. These could be filtered by annotation *body* or by *target* but these concepts will need to be communicated clearly in the user interface to users who have no prior knowledge of OA.
- Annotation motivation – to address the question, 'why did the author make the annotation?'

Considering the first bullet, classification of annotations by domain is an important capability and arguably a core part of the model. However, given the range of vocabularies available for classifying data - even within the Earth sciences – it may well require profiling for specific sub-domains or application areas. Some examples of vocabularies are:

- INSPIRE themes
- ISO Topic keywords
- SWEET Ontology
- GCMD
- GEMET

OA provides a means to do so via *tagging*. In this case, an annotation is set with the `oa:tagging` motivation. The body of the annotation contains the actual tag. A tag may be a simple text-based keyword or a semantic tag, a URI relating to a concept from a vocabulary. Using the OA semantic tags feature allows annotations can be classified with concepts from SKOS collections. Using the NERC Vocabulary Server[12] it is possible to map classification terms from one vocabulary to similar concepts in other vocabularies.

The process of tagging is itself an input for the faceted search, the tag property forming a category for search. The example below shows the classification of the ECMWF Operational Analyses data with the *Atmospheric conditions* concept from the INSPIRE themes.

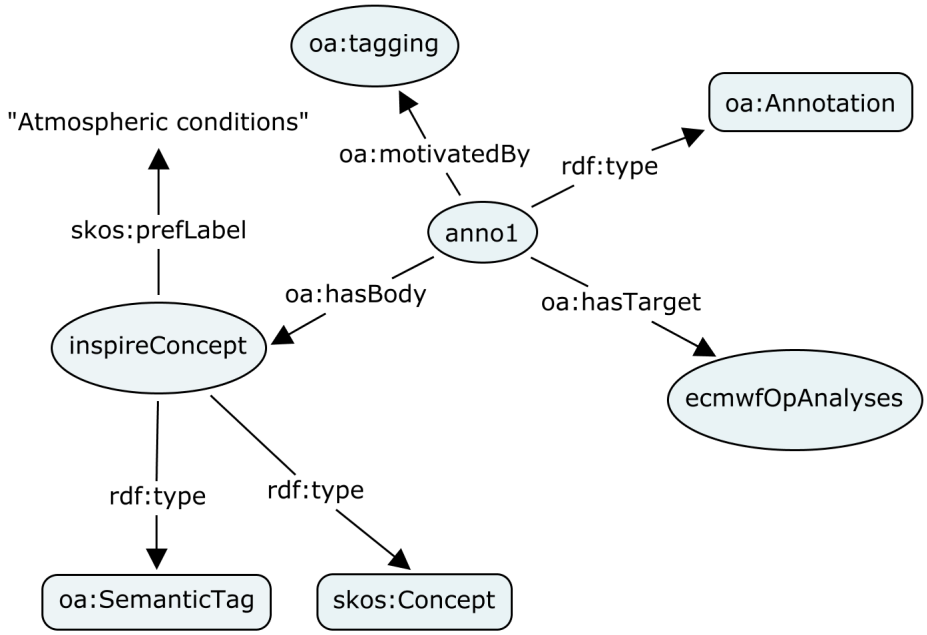

Figure 12: Tagging a Dataset with a SKOS Concept

```
@prefix oa: <http://www.openannotation.org/spec/core/> .
@prefix dctypes: <http://purl.org/dc/dcmitype/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<anno1> a oa:Annotation ;
       oa:hasBody <tag1> ;
       oa:hasTarget
       <http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_ECMWF-OP>
       ;
       oa:motivatedBy oa:tagging .

<http://vocab.ndg.nerc.ac.uk/term/P220/1/26> a oa:SemanticTag ;
<http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_ECMWF-OP> a
dctypes:Dataset .
```

Figure 13: Turtle - tagging a Dataset with a SKOS Concept

[12] http://vocab.nerc.ac.uk/

6 Guidelines for Integration with Existing Infrastructures

This section is intended to provide guidelines about the application of the model to existing infrastructures deployed at data provider sites. Data providers have their own existing data and metadata holdings. These each have their own structure with established models of usage and user communities associated.

OA is the basis is the basis for the CHARMe model, so we build on Linked Data principles, *resources used in annotations must be resolvable HTTP URIs*. This may require changes such as for example adding a RESTful interface to an existing repository to expose resources.

Resources themselves, may be structured or unstructured. If structured, an implementation decision needs to be made:

- Do nothing – the structured data is well understood and has supported software for parsing. It is sufficient, for the data to be referenced by CHARMe by URI. To CHARMe the structured data appears as *unstructured*.
- Create an RDF serialisation – this could be implemented as a façade over an existing interface. Representation of RDF enables greater potential for the linking of different resources together both internally and externally to a given repository. It also allows the mapping of relationships between data and potential for usage beyond what was originally intended. Careful consideration of the value of such work is needed weighed against the implementation effort required to make the changes. This in the scope of WP640 and WP700 tasks.