

Domain-specific modeling languages for Citizen Science data provenance

MDENet seedcorn fund project

Research Metadata

- FAIR Principles
 - Findability, Accessibility, Interoperability and Reuse
- Reusable datasets
 - Documenting the creation encourages use/reuse of datasets
 - User of a dataset is given an explanation of the creation process
- Citizen Science
 - Can be at a disadvantage lacking metadata expertise
 - Data produced is often mistrusted

Example Citizen Science project: <https://naturehood.uk/survey-your-space>

Approaches to lineage documentation



NONE - Do nothing!

- “My data explains itself!”
- Lack of metadata knowledge or training

Unstructured

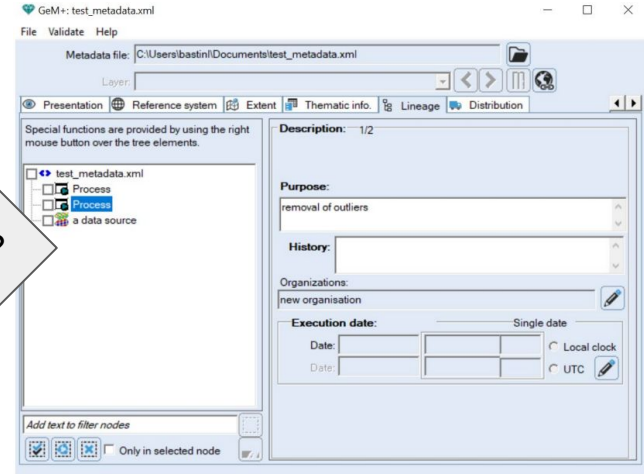
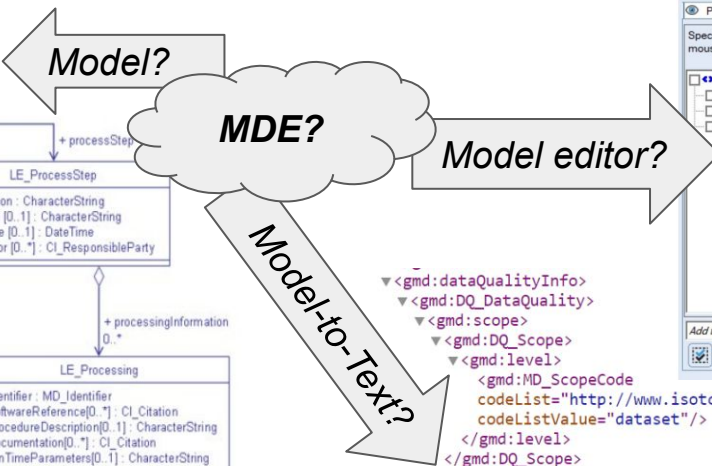
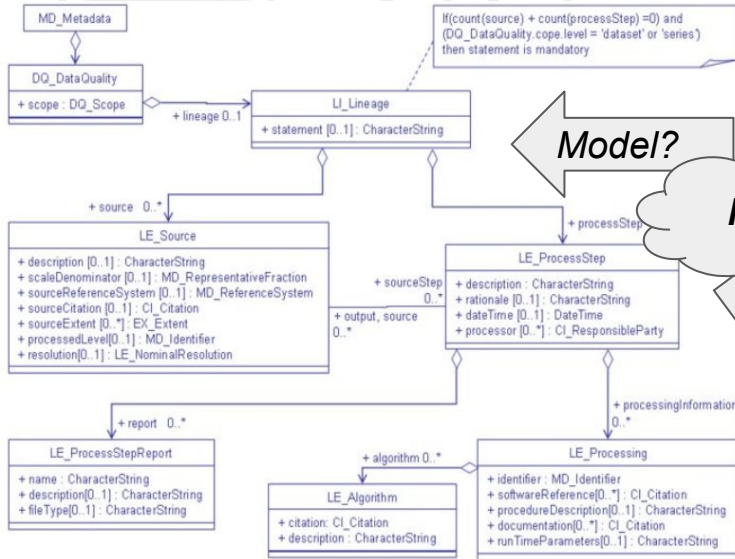
- Just write a document
- Not machine parsable
- May not follow a “standard”

Structured

- Lineage specifications
 - ISO19115-3
- Minimum level of detail
- Interchange formats (XML)
- Complex and difficult
 - Non-experts will struggle



ISO9115-3: UML, XML, Tools



```

<gmd:dataQualityInfo>
  <gmd:DQ_DataQuality>
    <gmd:scope>
      <gmd:DQ_Scope>
        <gmd:level>
          <gmd:MD_ScopeCode
            codeList="http://www.isotc211.org/2005/resources/codelist.xml#MD_ScopeCode"
            codeListValues="dataset"/>
        </gmd:level>
      </gmd:DQ_Scope>
    </gmd:scope>
  </gmd:dataQualityInfo>
  <gmd:lineage>
    <gmd:LI_Lineage>
      <gmd:statement>
        <gco:CharacterString>Due to the map generation method, the quality of the map can never be uniform. The overall quality of the map depends heavily on the individual quality of the data for the different countries.</gco:CharacterString>
      </gmd:statement>
    </gmd:LI_Lineage>
  </gmd:lineage>
</gmd:DQ_DataQuality>
</gmd:dataQualityInfo>
  
```

Model-driven solution

Aim: try and reduce the barrier of entry to creating lineage documentation that meets specifications like ISO19115-3.

- Modeling language (based on ISO19115-3)
 - Use generators for model editors and text file outputs
 - One example of file output: the standard XML files
- Assistive tools with a “text-like” editor (MPS, Abstract Syntax Tree)
 - Form-like structure, task automation, auto-complete, hints
- Model your project once, produce multiple deliverables

Project: MPS Prototype

Project participants:

- Aston University
 - Owen Reynolds, *PhD student*
 - Lucy Bastin,
Research metadata & Citizen Science expert
- University of York
 - Antonio Garcia-Dominguez, *MDE expert*
- Earthwatch
 - James Sprinks, *Citizen Science expert*

Why JetBrains MPS?

- Open-source language workbench
 - Structure, Editor, Type system
 - Transformations, Constraints
- Projectional editor
 - Quickly implements complex “text-like editor with a guide
 - Looks like a document but has cells
 - Intentions & auto-complete
 - Automated validation of user inputs

MPS: ISO 19115-3



```
AlienSpecies
├── Citations
├── Citizen_Photo
├── Contacts
├── Anon Citizen
├── Data processing department
├── Lineage Documents
└── Invasive Alien Species in Europe App

step Citizen identifies attributes
[LE_processStep]
Process step Citizen submits observation data to data management system
description [Having completed the data collection and identification the citizen uses the in app submit button.
             [Data is transmitted to the data management system
             [Citizen observation data requires further processing. (AI and expert validation.) ]
rationale
date and time <no dateTime>
reports
<< no reports >>
```



```
DQLineage
├── structure
│   ├── DQ_DataQuality
│   ├── LE_Algorithm
│   ├── LE_Processing
│   ├── LE_ProcessStep
│   ├── LE_ProcessStepReference
│   ├── LE_ProcessStepReport
│   ├── LE_Source
│   ├── LE_SourceReference
│   ├── LI_Lineage
│   ├── LineageDocument
│   └── List_ProcessStep
└── editor
    ├── DQ_DataQuality_Editor
    ├── LE_Algorithm_Editor
    ├── LE_Processing_Editor
    ├── LE_ProcessStep_Editor
    ├── LE_ProcessStepReference_Editor
    ├── LE_ProcessStepReport_Editor
    ├── LE_Source_Editor
    ├── LE_SourceReference_Editor
    ├── LI_Lineage_Editor
    └── LineageDocument_Editor

<default> editor for concept LE_ProcessStep
node cell layout:
[LE_processStep]
[>] Process step { name } <[
<- [/]
[>] description % description % <[
[>] rationale % rationale % <[
[>] date and time % dateTime % <[
[>] reports
[>] (/ % report % /) -1
[>] /empty cell: << no reports >>
[>] /folded cell: << ... reports ... >>
[>] /3
[>] processing information
[>] (/ % processingInformation % /) -1
[>] /empty cell: << no processing information >>
[>] /folded cell: << ... processing information ... >>
[>] /3
[>] Source inputs
[>] (/ % source % /) -1
[>] /empty cell: <default>
[>] /folded cell: << ... sources ... >>
[>] /3
```

Invasive Alien Species in Europe App — Konqueror

- Citizen identifies attributes**

This includes the provision of fact sheets, which describe the species covered by this particular mobile application and provide images as well as lists of commonly confused species. Element data is collected through an in app key pad

Inputs	Outputs
<ul style="list-style-type: none">citizenObservationPhoto	<ul style="list-style-type: none">citizenObservationAttributes
- Citizen submits observation data to data management system**

Having completed the data collection and identification the citizen uses the in app submit button. Data is transmitted to the data management system

Inputs	Outputs
<ul style="list-style-type: none">citizenObservationLocationcitizenObservationDateTimecitizenObservationPhotocitizenObservationAttributescitizenObservationComment	<ul style="list-style-type: none">citizenObservationDataSet



Next step: User testing

Earthwatch Citizen Science project managers

- What do Citizen Science project managers think of the modelling style approach to project documentation?
- Is the “text-like” editor interface helpful/intuitive?
- What other kinds of output formats would helpful?

Observations so far...

- Using ISO 19115-3 for documentation is hard
 - Overwhelming with the details
 - Unclear documentation, which is also overwhelmingly big!
- Model representations can be tuned to audiences
 - One language model and one project model
 - Multiple editors and outputs
- Modelling tools and lineage specifications share a problem
 - Both become complicated with the levels of detail involved
 - Higher levels of abstraction and assistive tools help

Future work

- User feedback will guide
 - Additional output types (graphical representations of process flows?)
 - Editor experience (different projectional editors)
- Tool delivery via the web
 - “Cloud based” project modelling tool
 - Remove the requirements for installing software
 - Enable collaborative modelling for project documentation
- Follow up research grant application
 - To build a set of standards-based notations for the Citizen Science community using MDE tools to support the development process



Thank you for attending!

Any questions?

Contact details:

Antonio Garcia-Dominguez

a.garcia-dominguez@york.ac.uk

Lucy Bastin

l.bastin@aston.ac.uk

Owen Reynolds

180200041@aston.ac.uk