Unleashing the Potential: How Automated Metadata Creation Can Improve Your Data Quality and Accessibility

Elena Robu | Astun Technology
What is Metadata anyway?
What is Metadata anyway?
What is Metadata anyway?

In simple terms, metadata is data about data.
What is Q-FAIR all about?
What is **Q-FAIR** all about?
What is Q-FAIR all about?

- Quality
- is just a hyphen
What is **Q-FAIR** all about?

- **Quality**
- is just a hyphen
- **Findable**
What is Q-FAIR all about?

- Quality
- Findable
- - is just a hyphen
- Accessible
What is Q-FAIR all about?

- Quality
- Findable
- Accessible
- Interoperable
What is Q-FAIR all about?

Quality
Findable
Interoperable

- is just a hyphen
Accessible
Reusable
Background
Background

Why are we doing this?

🤔
Background

How do we do this now? 🕒
Background

How do we do this now? 🕒

Metadata Crawler open source plugin for Talend Open Studio (aka Crawler)
What are these fields?
What are these fields?

- Abstract
- Alternative Title
- Bounding box
- Constraints (usage and licensing)
- Contact Details
- Date
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
What are these fields?

Abstract
Alternative Title
Bounding box
Constraints (usage and licensing)
Contact Details
Date
Geographic Extent (keyword)
INSPIRE Keywords (controlled)

Keywords (free text or controlled)
Lineage
Spatial Reference System
Resolution
Temporal Extent
Title
Topic Category (controlled)
Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Bounding box
- Constraints (usage and licensing)
- Contact Details
- Date
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Lineage
- Spatial Reference System
- Resolution
- Temporal Extent
- Topic Category (controlled)
- Title
- Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Constraints (usage and licensing)
- Contact Details
- Date
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Lineage
- Spatial Reference System
- Resolution
- Temporal Extent
- Title
- Topic Category (controlled)
- Update Frequency
<table>
<thead>
<tr>
<th>What are these fields?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
</tr>
<tr>
<td>Alternative Title</td>
</tr>
<tr>
<td>Constraints (usage and licensing)</td>
</tr>
<tr>
<td>Contact Details</td>
</tr>
<tr>
<td>Geographic Extent (keyword)</td>
</tr>
<tr>
<td>INSPIRE Keywords (controlled)</td>
</tr>
<tr>
<td>Keywords (free text or controlled)</td>
</tr>
<tr>
<td>Lineage</td>
</tr>
<tr>
<td>Spatial Reference System</td>
</tr>
<tr>
<td>Resolution</td>
</tr>
<tr>
<td>Temporal Extent</td>
</tr>
<tr>
<td>Title</td>
</tr>
<tr>
<td>Topic Category (controlled)</td>
</tr>
<tr>
<td>Update Frequency</td>
</tr>
</tbody>
</table>
What are these fields?

- Abstract
- Alternative Title
- Constraints (usage and licensing)
- Contact Details
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Lineage
- Resolution
- Temporal Extent
- Title
- Topic Category (controlled)
- Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Constraints (usage and licensing)
- Contact Details
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Lineage
- Resolution
- Temporal Extent
- Topic Category (controlled)
- Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Constraints (usage and licensing)
- Contact Details
- Keywords (free text or controlled)
- Lineage
- Resolution
- Temporal Extent
- Topic Category (controlled)
- Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Contact Details
- Keywords (free text or controlled)
- Lineage
- Resolution
- Temporal Extent
- Topic Category (controlled)
- Update Frequency
What are these fields?

Abstract
Alternative Title
Geographic Extent (keyword)
INSPIRE Keywords (controlled)

Keywords (free text or controlled)
Lineage
Resolution
Temporal Extent
Topic Category (controlled)
Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Resolution
- Temporal Extent
- Topic Category (controlled)
- Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Temporal Extent
- Topic Category (controlled)
- Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Topic Category (controlled)
- Update Frequency
What are these fields?

- Abstract
- Alternative Title
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Topic Category (controlled)
What are these fields?

- Abstract
- Alternative Title
- Geographic Extent (keyword)
- INSPIRE Keywords (controlled)
- Keywords (free text or controlled)
- Topic Category (controlled)
How do we automate these?
How do we automate these?

spaCy
How do we automate these?

"industrial strength open source NLP in Python"
Geographic Keywords
Geographic Keywords

NER model trained on OS Open Names
Geographic Keywords

NER model trained on OS Open Names

Pattern recognition model (postcodes)
Non-Geographic Keywords
Non-Geographic Keywords

pytextrank
Non-Geographic Keywords

pytextrank

KeyBERT
Non-Geographic Keywords
INSPIRE Keywords and Topic Categories
INSPIRE Keywords and Topic Categories

Text Classification
Everything Else
Everything Else

USE CHATGPT

FOR ALL THE THINGS!!
How will this work IRL?
<table>
<thead>
<tr>
<th>Date</th>
<th>Organisation</th>
<th>Inspire_keyword</th>
<th>Topic_category</th>
<th>Geographic area</th>
<th>Keywords</th>
<th>Record abstract</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>23-07-2013</td>
<td>Astun Technology</td>
<td>Transport networks</td>
<td>society</td>
<td>Aberdeen, Scotland</td>
<td>Aberdeen, Cults</td>
<td>On July 23, 2013, Astun Technology created a descriptive dataset on transport networks in Aberdeen, Scotland. The dataset focuses on the smallest geographic area that includes all locations in Aberdeen, including Cults and Tree.</td>
<td>Transport Networks in Aberdeen, Scotland: Cults and Tree</td>
</tr>
<tr>
<td>23-07-2013</td>
<td>Astun Technology</td>
<td>Protected sites</td>
<td>society</td>
<td>United Kingdom</td>
<td>Historic County</td>
<td>This dataset provides information on protected sites in the United Kingdom, specifically focusing on East Lothian Historic County. Created on 23-07-2013, it includes data on the smallest geographic area that encompasses all the listed locations. The dataset also incorporates information on historic counties and boundary lines.</td>
<td>Protected Sites in the United Kingdom's East Lothian Historic County</td>
</tr>
</tbody>
</table>
Transport Networks in Aberdeen, Scotland: Cults and Tree

Also named: Tree_Preservation_Orders
On July 23, 2013, Astun Technology created a descriptive dataset on transport networks in Aberdeen, Scotland. The dataset focuses on the smallest geographic area that includes all locations in Aberdeen, including Cults and Tree.

Discover data

Download

Description

This dataset is published in the database "data/crawlerinput/Tree_Preservation_Orders.shp" in table "Description".

Format

ESRI Shapefile
Questions?
Thank you!

@AlexaSmile15  elenarobu
elenarobu@astuntechnology.com
astuntechnology.com/contact-us