

PIDs at LINDAT/CLARIN (Czech CLARIN)

Pavel Straňák



LINDAT/CLARIN repository

- Dspace: uses handles internally. Cannot work without a handle server
 - Even if we use DOIs, they must be used together with our handles anyway
- Record = handle
 - No handles for bitstreams
 - standardised URL structure (a suffix) + REST API for bitstream access
 - Bitstream URLs in metadata
 - `<cmd:ResourceType>Resource</cmd:ResourceType>`

The screenshot shows a digital repository record view for English morphological models. The main content area includes:

- Language(s):** English
- Description:** English models for MorphoDiTa, providing morphological analysis, morphological generation and part-of-speech tagging. The morphological dictionary is created from Morphium and SCOWL (Spell Checker Oriented Word Lists), the PoS tagger is trained on WSJ (Wall Street Journal).
- Publisher:** Charles University in Prague, UFAL
- Subject(s):** MorphoDiTa, English, morphological analysis, morphological generation, PoS tagging
- Files in this item:** Two zip files are listed:
 - english-morphium-wsj-140407.zip** (13.12 MB, application/zip): English Models (Morphium + WSJ) for MorphoDiTa. Minor fix to english-morphium-wsj-140304 allowing to recognize "non-" as a negative prefix in addition to "non".
 - english-morphium-wsj-140304.zip** (13.12 MB, application/zip): English Models (Morphium + WSJ) for MorphoDiTa

A license notice states: "This item is **Publicly Available** and licensed under: Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)".

The right sidebar contains navigation and administrative options:

- Profile
- Submissions
- Context
 - Edit this item
 - Export Item
 - Export Metadata
- Administrative
 - Access Control
 - Registries
 - Items
 - Collections & Communities
 - Withdrawn Items
 - Control Panel
 - Statistics
 - Import Metadata
 - Curation Tasks
 - Licenses
 - Handles
- General Information
 - Deposit

Record View – resources

(CMDI equivalent on the next slide)

The screenshot shows a digital repository record view. The main content area includes:

- Language(s):** English
- Description:** English models for MorphoDiTa, providing morphological analysis, morphological generation and part-of-speech tagging. The morphological dictionary is created from Morphium and SCOWL (Spell Checker Oriented Word Lists), the PoS tagger is trained on WSJ (Wall Street Journal).
- Publisher:** Charles University in Prague, UFAL
- Subject(s):** MorphoDiTa, English, morphological analysis, morphological generation, PoS tagging
- Files in this item:** A list of two files, each 13.12 MB, application/zip format. The first file is named 'english-morphium-wsj-140407.zip' and its description is 'English Models (Morphium + WSJ) for MorphoDiTa. Minor fix to english-morphium-wsj-140304 allowing to recognize "non-" as a negative prefix in addition to "non"'. A red circle highlights the 'Download file' button for this file. The second file is named 'english-morphium-wsj-140304.zip' and its description is 'English Models (Morphium + WSJ) for MorphoDiTa'. A 'Download file' button is also present for this file.

On the right side, there is a sidebar with navigation options: Profile, Submissions, Context (Edit this item, Export Item, Export Metadata), Administrative (Access Control, Registries, Items, Collections & Communities, Withdrawn Items, Control Panel, Statistics, Import Metadata, Curation Tasks, Licenses, Handles), and General Information (Deposit).

Record View – resources

(CMDI equivalent on the next slide)

The screenshot shows a digital repository record view. The main content area includes:

- Language(s):** English
- Description:** English models for MorphoDiTa, providing morphological analysis, morphological generation and part-of-speech tagging. The morphological dictionary is created from Morphium and SCOWL (Spell Checker Oriented Word Lists), the PoS tagger is trained on WSJ (Wall Street Journal).
- Publisher:** Charles University in Prague, UFAL
- Subject(s):** MorphoDiTa, English, morphological analysis, morphological generation, PoS tagging
- Files in this item:** A list of two files, each with a "Download file" button circled in red:
 - File 1:** Name: english-morphium-wsj-140407.zip, Size: 13.12 MB, Format: application/zip, Description: English Models (Morphium + WSJ) for MorphoDiTa. Minor fix to english-morphium-wsj-140304 allowing to recognize "non-" as a negative prefix in addition to "non".
 - File 2:** Name: english-morphium-wsj-140304.zip, Size: 13.12 MB, Format: application/zip, Description: English Models (Morphium + WSJ) for MorphoDiTa

A light blue banner indicates the item is "Publicly Available" and licensed under Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0).

The right sidebar contains navigation and administrative options:

- Profile
- Submissions
- Context
 - Edit this item
 - Export Item
 - Export Metadata
- Administrative
 - Access Control
 - Registries
 - Items
 - Collections & Communities
 - Withdrawn Items
 - Control Panel
 - Statistics
 - Import Metadata
 - Curation Tasks
 - Licenses
 - Handles
- General Information
 - Deposit

Record View – resources

(CMDI equivalent on the next slide)

Language(s) English

Description English models for MorphoDiTa, providing morphological analysis, morphological generation and part-of-speech tagging.

The morphological dictionary is created from Morphium and SCOWL (Spell Checker Oriented Word Lists), the PoS tagger is trained on WSJ (Wall Street Journal).

Publisher Charles University in Prague, UFAL

Subject(s) MorphoDiTa English morphological analysis morphological generation PoS tagging

Show full item record

Files in this item [Download all files in item \(26.24 MB\)](#)

This item is **Publicly Available** and licensed under:
Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)

Name english-morphium-wsj-140407.zip
Size 13.12 MB
Format application/zip
Description English Models (Morphium + WSJ) for MorphoDiTa. Minor fix to english-morphium-wsj-140304 allowing to recognize "non-" as a negative prefix in addition to "non". [Download file](#)

Name english-morphium-wsj-140304.zip
Size 13.12 MB
Format application/zip
Description English Models (Morphium + WSJ) for MorphoDiTa [Download file](#)

Administrative

- Access Control
- Registries
- Items
- Collections & Communities
- Withdrawn Items
- Control Panel
- Statistics
- Import Metadata
- Curation Tasks
- Licenses
- Handles

General Information

- Deposit

Record View – resources

(CMDI equivalent on the next slide)

```
<cmd:ResourceProxyList>
<cmd:ResourceProxy id="lp_242">
<cmd:ResourceType>LandingPage</cmd:ResourceType>
<cmd:ResourceRef>
http://hdl.handle.net/11858/00-097C-0000-0023-68D9-0
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="uri_1">
<cmd:ResourceType mimetype="text/html">Resource</cmd:ResourceType>
<cmd:ResourceRef>
http://ufal.mff.cuni.cz/morphodita/users-manual#english-morphium-wsj
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="_766">
<cmd:ResourceType mimetype="application/zip">Resource</cmd:ResourceType>
<cmd:ResourceRef>
https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/
11858/00-097C-0000-0023-68D9-0/english-morphium-wsj-140407.zip?sequence=3
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="_761">
<cmd:ResourceType mimetype="application/zip">Resource</cmd:ResourceType>
<cmd:ResourceRef>
https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/
11858/00-097C-0000-0023-68D9-0/english-morphium-wsj-140304.zip?sequence=1
</cmd:ResourceRef>
</cmd:ResourceProxy>
</cmd:ResourceProxyList>
```

```
<cmd:ResourceProxyList>
<cmd:ResourceProxy id="lp_242">
<cmd:ResourceType>LandingPage</cmd:ResourceType>
<cmd:ResourceRef>
http://hdl.handle.net/11858/00-097C-0000-0023-68D9-0
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="uri_1">
<cmd:ResourceType mimetype="text/html">Resource</cmd:ResourceType>
<cmd:ResourceRef>
http://ufal.mff.cuni.cz/morphodita/users-manual#english-morphium-wsj
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="_766">
<cmd:ResourceType mimetype="application/zip">Resource</cmd:ResourceType>
<cmd:ResourceRef>
https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/
11858/00-097C-0000-0023-68D9-0/english-morphium-wsj-140407.zip?sequence=3
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="_761">
<cmd:ResourceType mimetype="application/zip">Resource</cmd:ResourceType>
<cmd:ResourceRef>
https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/
11858/00-097C-0000-0023-68D9-0/english-morphium-wsj-140304.zip?sequence=1
</cmd:ResourceRef>
</cmd:ResourceProxy>
</cmd:ResourceProxyList>
```



```
<cmd:ResourceProxyList>
<cmd:ResourceProxy id="lp_242">
<cmd:ResourceType>LandingPage</cmd:ResourceType>
<cmd:ResourceRef>
http://hdl.handle.net/11858/00-097C-0000-0023-68D9-0
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="uri_1">
<cmd:ResourceType mimetype="text/html">Resource</cmd:ResourceType>
<cmd:ResourceRef>
http://ufal.mff.cuni.cz/morphodita/users-manual#english-morphium-wsj
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="_766">
<cmd:ResourceType mimetype="application/zip">Resource</cmd:ResourceType>
<cmd:ResourceRef>
https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/
11858/00-097C-0000-0023-68D9-0/english-morphium-wsj-140407.zip?sequence=3
</cmd:ResourceRef>
</cmd:ResourceProxy>
<cmd:ResourceProxy id="_761">
<cmd:ResourceType mimetype="application/zip">Resource</cmd:ResourceType>
<cmd:ResourceRef>
https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/
11858/00-097C-0000-0023-68D9-0/english-morphium-wsj-140304.zip?sequence=1
</cmd:ResourceRef>
</cmd:ResourceProxy>
</cmd:ResourceProxyList>
```

Granularity of PIDs

- Typical: a corpus, individual files unimportant, with limited to no MD
 - One PID in repository, part identifiers in other apps (interfaces) – subsets
- Sometimes: rich metadata for files, still lot of files (e.g. recording + transcripts + metadata on speakers)
 - Repository is NOT a good place to work with this rich metadata and data
 - In repository still one PID by default. subset PIDs possible, but should resolve into views in a specialised interface. E.g. a corpus search engine and browser.
- New records for explicit subsets always possible (*dc.relation.ispartof*)

Vyhledat slovo nebo řetěz slov:

Slovo v levém kontextu:

Rozsah levého kontextu (max. 25):

Vyhledat

Lemma 1. slova:

Tag 1. slova:

Slovo v pravém kontextu:

Rozsah pravého kontextu (max. 25):

Mluvčí:

respondent žena
respondent muž

Typ školy: Ročník školy:

První jazyk: Doma mluví:

Mluví rómsky: Sociálně vyloučená komunita:

Nahrávka

Věk respondenta od do

Popis tagu 1. slova

Pozice 1 - Slovní druh

Pozice 2 - Detailní určení slovi

Pozice 3 - Jmenný rod (GEND

Pozice 4 - Číslo (NUMBER)

Pozice 5 - Pád (CASE)

Pozice 6 - Přivlastňovací rod (

Pozice 7 - Přivlastňovací číslo

Pozice 8 - Osoba (PERSON)

Pozice 9 - Čas (TENSE)

Pozice 10 - Stupeň (GRADE)

Pozice 11 - Negace (NEGATIC

Pozice 12 - Aktivum/pasívum

Pozice 13 - Nepoužito (RESEF

Pozice 14 - Nepoužito (RESEF

Complex dataset view

A proper way to search complex metadata

DataCite usage?

- It is possible to get DOIs from DataCite
- Czechia has no member, but German members are willing to register our data
- Good: We get DOIs. DOIs are well known and sometimes wanted (ministries, Thompson Reuters (i.e. IF, i.e. ministries ...))
- Bad: We are not owners of our PIDs
- Ugly: metadata needed for DataCite is not direct subset of our metadata
- Why DOIs? Handles can do the same, just not so known.
 - Is this a reason to resign, or rather promote (generic) handles?

DOI vs. (generic) Handle

- DOIs impose some limitations on handles
 - Very reasonable limitations, especially for citation purposes
- Clarin requires handles, they can be DOIs
 - Clarin imposes some limitations too
 - They are very similar. E.g. all handles must have some metadata, human readable on resolving the handle via web browser
- DOIs are well established:
 - Many applications are tuned for them (CrossRef, ORCID, etc.)
 - Many use cases expect them (publishers, funders, citation indexes)

We have 2 options

- Require DOIs:
 - handles, popular, free, no real drawbacks
 - usefull services like ORCID linking, Zenodo-GitHub publishing, **ODIN, CrossRef**
- Promote generic handles (increase their usefulness):
 - **No service above really needs DOIs**
 - handles of good quality (basic metadata) should be allowed too. We have to work on that.
 - Not having unified minimal metadata might be the core problem.

We have decided to promote handles (at least for now)

- Thompson-Reuters DataCitation Index
 - We have applied, application accepted
 - Metadata checked and found OK
 - We should appear in the index ...
- We will see if some situation requires DOIs