Raznoliki metapodatkovni standardi in podatkovni formati v humanistiki: izzivi in rešitve pri vzpostavljanju certificiranega podatkovnega središča za digitalno humanistiko

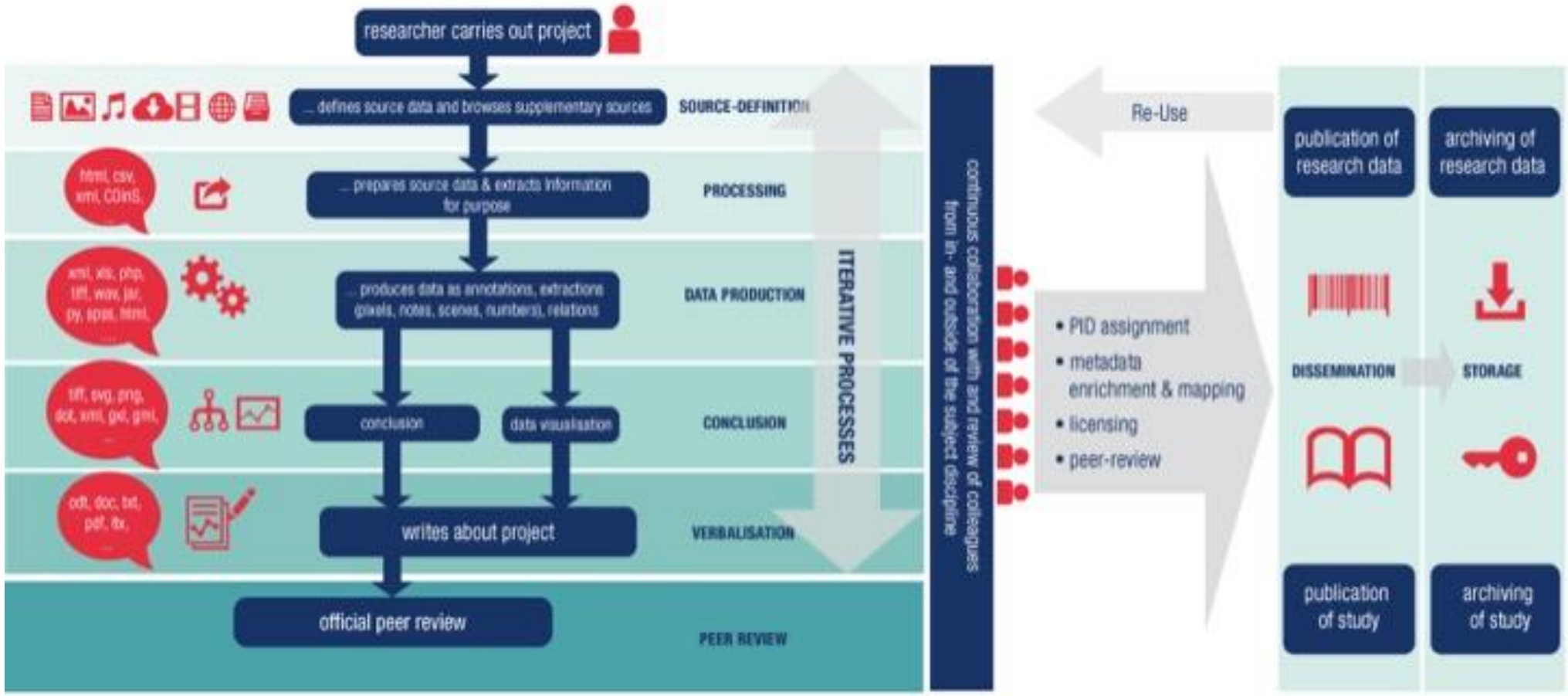
Andrej Pančur
Inštitut za novejšo zgodovino, DARIAH-SI
andrej.pancur@inz.si

Konferenca Odprti raziskovalni podatki v Sloveniji
Maribor, 14. 11. 2019

Kaj je DARIAH

- DARIAH je evropska ESFRI raziskovalna infrastruktura za raziskovalce iz umetnosti in humanistike, ki pri svojem delu uporabljajo računalniške metode.
- DARIAH je mreža ljudi, strokovnega znanja, informacij, znanja, podatkov, metod, orodij in tehnologij.
- DARIAH podpira digitalno raziskovanje in poučevanje digitalno podprtih raziskovalnih metod.
- DARIAH ohranja, zagotavlja dostop in diseminacijo rezultatov digitalno humanističnih raziskav.
- DARIAH spodbuja uporabo najboljših raziskovalnih praks, metod in tehničnih standardov.

DARIAH ni zgolj podatkovno središče za raziskovalne podatke v skladu z mednarodnimi standardi, temveč pokriva celoten krog raziskovalnih podatkov v digitalni humanistiki



Naloge podatkovnih središč v humanistiki: skrb za raziskovalne podatke v najširšem pomenu besede

- Skrb za podatke po zaključku običajnega življenjskega cikla podatkov.
- Skrb za predstavitveno okolje in aplikacije, ki omogočajo interpretacijo podatkov, iskanje, filtriranje in brskanje po podatkih ter njihovo povezovanje. Prezentacija podatkov kot sestavni del znanstvene argumentacije.
- Skrb za programsko kodo, na kateri temelji prezentacija in
- skrb za akademske programe, ki so sestavni del znanstvene argumentacije v digitalni humanistiki.

Kaj so raziskovalni podatki v digitalni humanistiki?

- Raziskovalni podatki v digitalni humanistiki so vsi viri in rezultati, ki se jih zbira, opisuje, vrednoti in/ali proizvaja v kontekstu raziskav v umetnosti in humanistiki in katere se lahko (dolgoročno) hrani v strojno berljivi obliki za namen arhiviranja, citiranosti in nadaljne uporabe. (DARIAH-DE)
- Naravoslovje in družboslovje: večinoma podatki iz meritev, vprašalnikov ipd.
- Humanistika: večinoma uporaba kulturnih objektov kot so rokopisi, besedila, slike, posnetki ipd., ki se jih kot digitalne surogate nato lahko še dodatno obdela, vizualizira, označi, poveže in interpretira.

Raznolike družboslovne in humanistične vede in skupnosti = raznoliki podatki

Uporabljani metapodatkovni standardi:

- CIDOC-CRM, CMDI, DataCite, DCDDM, EAD, DDI, Dublin Core, EDM, FITS, IIIF, METS, MODS, ODRL, OLAC, PREMIS, QuDEx, teiHeader, itd.

Uporabljani podatkovni formati:

- CHAT, CSV, FOLIA, EAF, EXCEL, EXMARaLDA, JPEH, SAS, SPSS, STATA, TAR, TCF, TEI, TIFF, Triple-S, TSV, ZIP, itd.

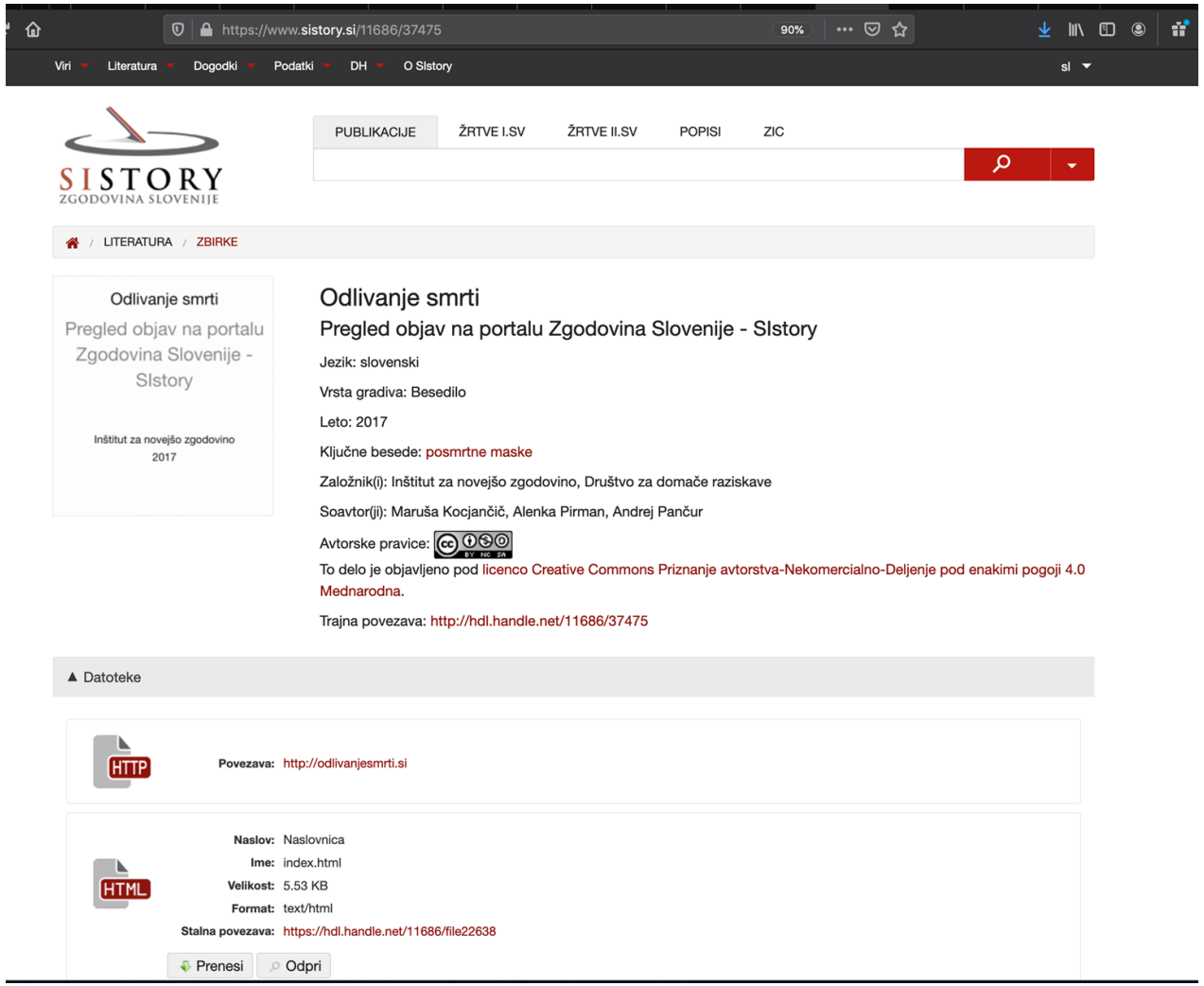
V humanistiki bi podatkovni centri morali upoštevati

- Najbolj pogosti metapodatkovni standardi v DARIAH skupnosti:
  - TEI (Text Encoding Initiative): teiHeader
  - CIDOC Conceptual Reference Model (CRM)
  - Dublin Core
- Zaradi potrebe po interoperabilnosti še: DataCite oz. OpenAIRE, EDM (Europeana Data Model), IIIF (International Image Interoperability Framework)
- DARIAH-SI vsaj še: EAD (Encoded Archival Description), LIDO (Lightweight Information Describing Objects), METS (Metadata Encoding and Transmission Standard), MODS (Metadata Object Description Schema), PREMIS (Preservation Metadata Implementation Strategies)
- Najbolj pogosti podatkovni formati: TEI, TIFF, JPEG

Kaj smo imeli na razpolago

Portal Zgodovina Slovenije – SIstory

- Portal kot repozitorij, digitalna knjižnica, dodatne baze podatkov
- HTML5, CSS, MySQL, PHP, ElasticSearch, Handle strežnik, OAI-PMH, API
- Dublin Core aplikacijski profil



Pri iskanju rešitve smo izhajali iz sledečih načel

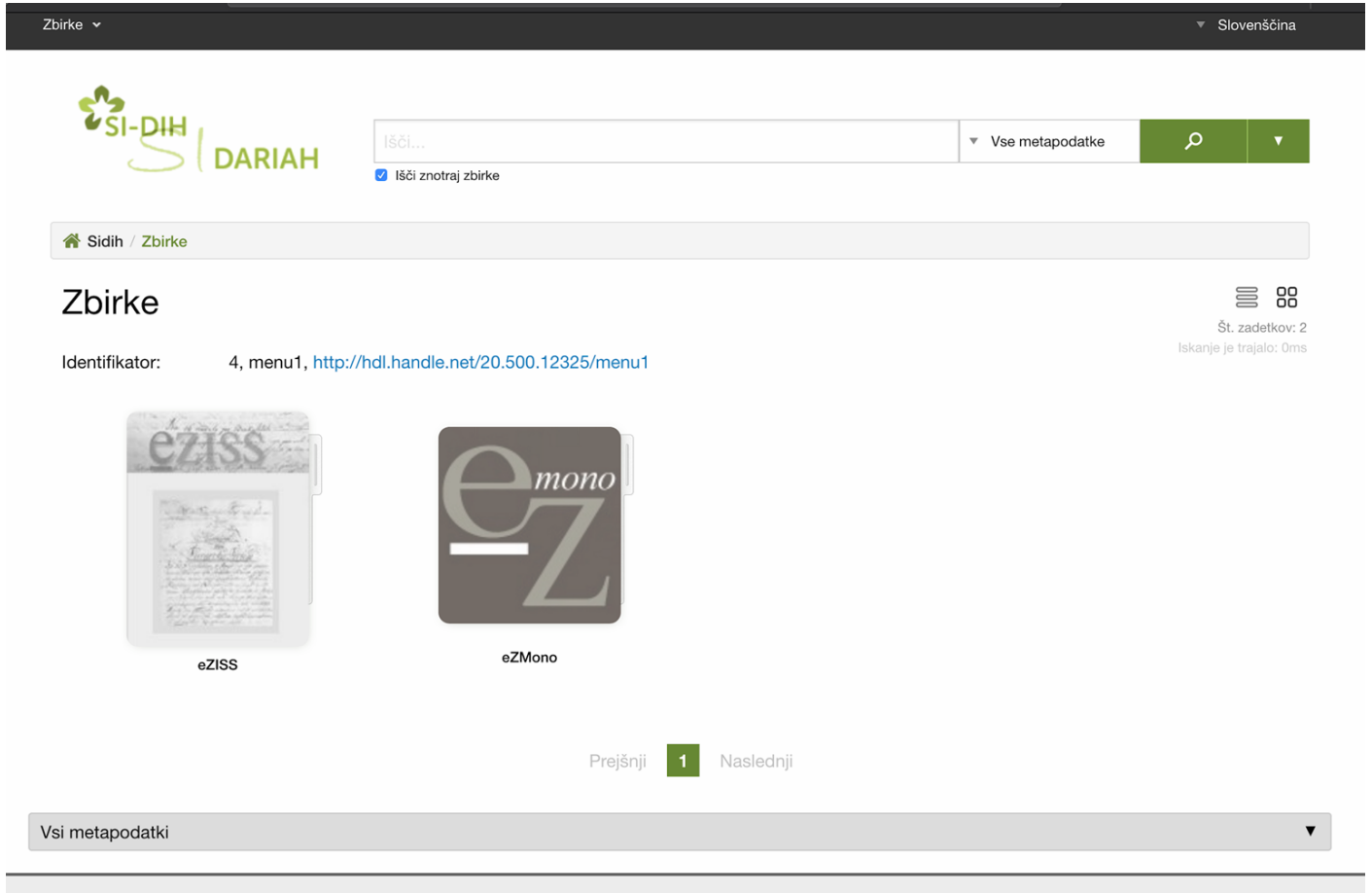
- **Enostavnost**: uporaba uveljavljenih, preizkušenih in zelo razširjenih tehnologij in standardov (npr. HTML, CSS, PHP, JavaScript, MySQL, ElasticSearch, RESTful, Web API), ki jih zunanji izvajalci dobro poznajo.
- **Poznavanje**: uporaba tehnologij in standardov, ki jih sami dobro poznamo in obvladamo (npr. XML tehnologije, metapodatkovni standardi s področja humanistike in umetnosti).
- **Fleksibilnost**: Fleksibilno in modularno nadgrajevanje obstoječih tehnologij v skladu z novimi znanji in spoznanji (mdr. semantični splet, povezani odprti podatki).
- **Odprtost**: uporaba odprtih (nelastniških) standardov: odprta koda in odprti podatki.

Pri iskanju rešitve smo preizkusili nekaj obstoječih odprtokodnih platform za upravljanje z digitalnimi objekti

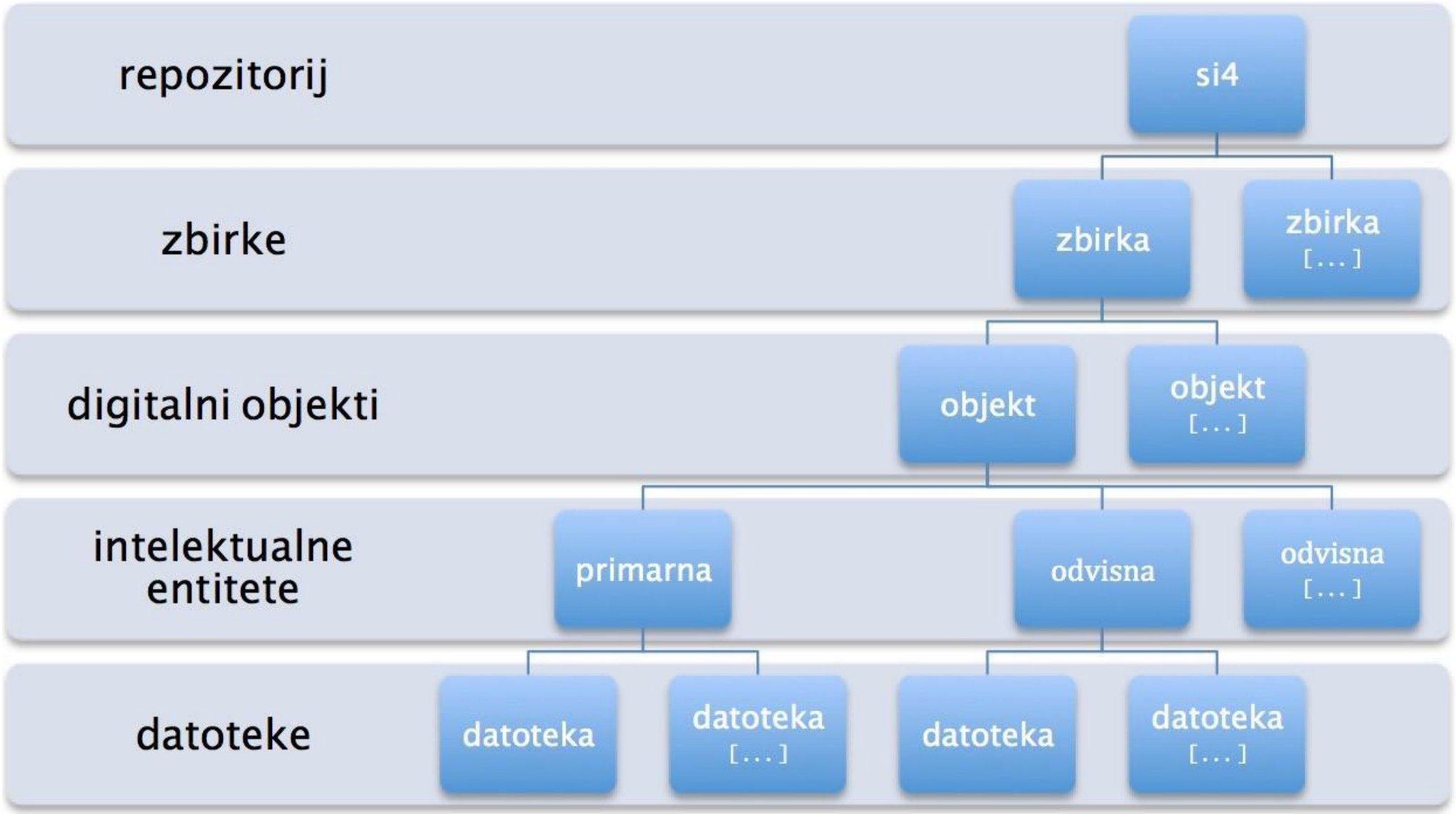
- Repozitoriji, ki bi jih lahko enostavno implementirali (AtoM, Omeka), nimajo zahtevanih funkcionalnosti za upravljanje raznolikih metapodatkovnih standardov in podatkovnih formatov.
- Repozitoriji, ki so bili dovolj fleksibilni, da jih lahko povsem prilagodimo našim potrebam (npr. Fedora Commons), pa so po drugi strani zelo zahtevni za ustrezno implementacijo. Njihova prilagoditev zahteva različna specialna znanja, ki med programerji niso splošno razširjena.
- Preizkusili smo Java aplikacijo Cirilo za upravljanje Fedora Commons repozitorijev (razvito za namen https://gams.uni-graz.at/):
  - Prednosti: predvsem TEI in LIDO; primerno za naše potrebe; odprtokodno; uspešna postavitev testne verzije.
  - Slabosti: pri razvoju novih funkcionalnosti in podatkovnih modelov, ki bi ustrezale našim specifilnim potrebam spet prejšnja težava; temeljila na zastareli (3.7) verziji repozitorija Fedora Commons; bila v postopku kompleksne nadgradnje
- **Rešitev**: razvoj lastne odprtokodne infrastrukture, ki sledi prej navedenim načelom (enostavnost, poznavanje, fleksibilnost, odprtost) => Preprosta raziskovalna infrastruktura za kompleksne raziskovalne podatke v humanistiki - **si4** (**S**imple research **I**nfrastructure FOR complex research data in digital humanities)

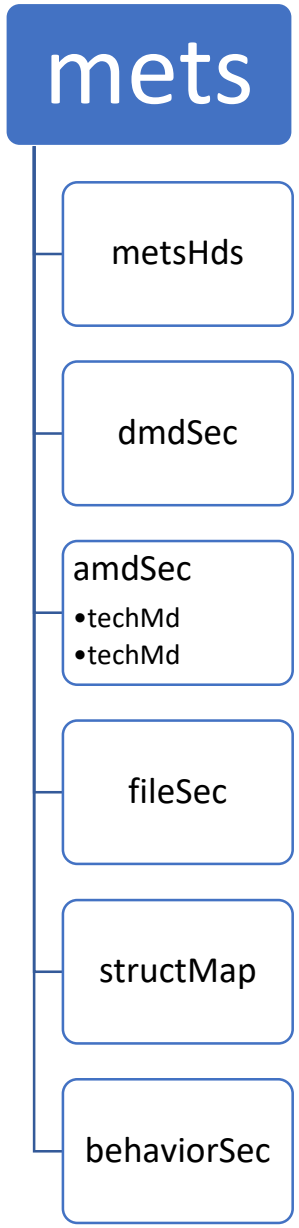
si4 repozitorij implementirali v nov portal **SIstory** (namenjen kulturni dediščini, začeli dolgotrajno fazo selitve) in nov repozitorij **SI-DIH** za slovensko digitalno humanistiko https://sidih.si/

- HTML5, CSS, PHP
- MySQL (relacije, identifikatorji, XML)
- ElasticSearch (JSON = metapodatki, indeksacija)
- Appache Tika
- MD5 Checksum
- Handle strežnik
- IIIF strežnik
- OAI-PMH



Izkušnje in rešitve za SIstory => vsebiski model repozitorija si4

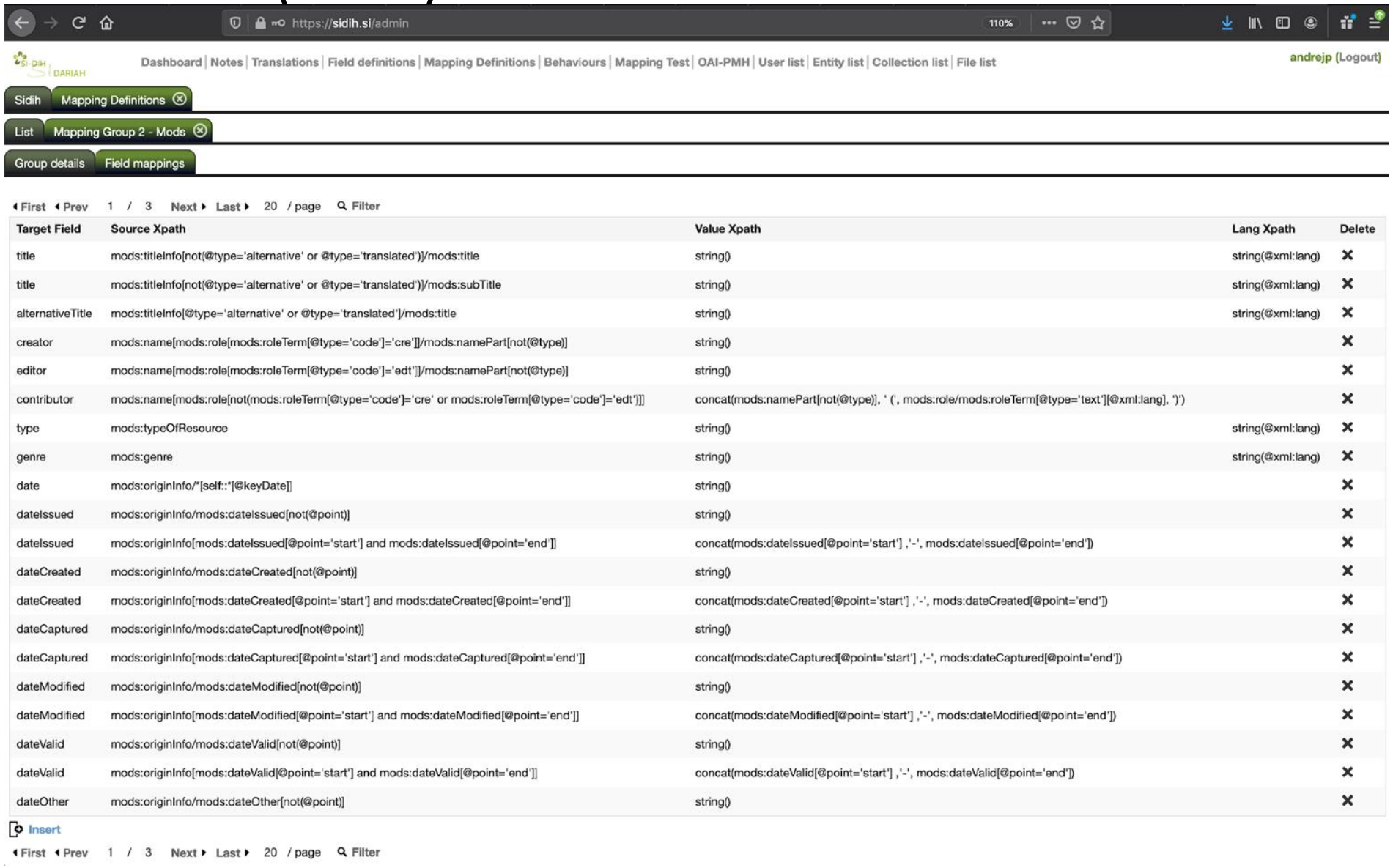




si4 METS profil

- Opisni metapodatki (dmdSec): Karkoli ti srce poželi ;-) , če je veljaven XML format. Načeloma pa v humanistiki običajni metapodatkovni standardi!
- amdSec/techMd[2]: dodatni tehnični metapodatki: umaknjena entiteta, nova verzija digitalnega objeka, dodaten opis vsebine v poljubnem HTML, zunanja zbirka, PDF stran.
- Vse ostalo ima fiksno strukturo in vsebino.

XPath mapiranje iz METS opisnih metapodatkov (XML) v ElasticSearch (JSON)



Mapiranje iz testnih MODS in DC opisnih metapodatkov (XML) v ElasticSearch (JSON)

Result Si4 | XPath Evaluate Test

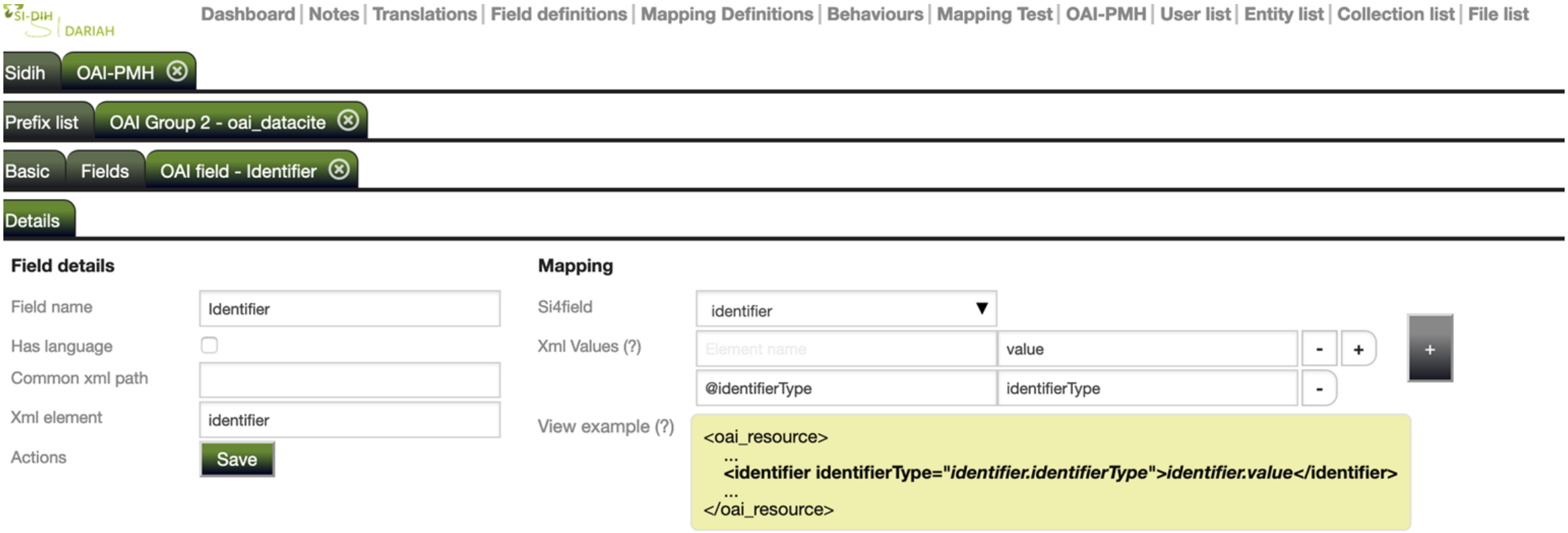
t Mets Xml

```xml
<?xml version="1.0" encoding="UTF-8"?>

<!-- This is a test xml -->

<METS:mets xmlns:METS="http://www.loc.gov/METS/"
           xmlns:xlink="http://www.w3.org/TR/xlink"
           xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
           TYPE="entity"
           ID="si4.test123"
           OBJID="http://hdl.handle.net/20.500.12325/test123">
    <METS:metsHdr CREATEDATE="2019-11-12T14:03:10" LASTMODDATE="2019-11-12T14:03:10" RECORDS
        <METS:agent ROLE="DISSEMINATOR" TYPE="ORGANIZATION">
            <METS:name>Sidih</METS:name>
            <METS:note>https://sidih.si</METS:note>
        </METS:agent>
        <METS:agent ROLE="CREATOR" ID="3" TYPE="INDIVIDUAL">
            <METS:name>Pančur, Andrej</METS:name>
        </METS:agent>
    </METS:metsHdr>
    <METS:dmdSec ID="description">
        <METS:mdWrap MDTYPE="DC">
            <METS:xmlData xmlns:dc="http://purl.org/dc/elements/1.1/"
                          xmlns:dcterms="http://purl.org/dc/terms/"
                          xmlns:dcmitype="http://purl.org/dc/dcmitype/">
                <dc:title xml:lang="slv">Test DC slv title</dc:title>
                <dc:title xml:lang="slv">Test DC slv title 2</dc:title>
                <dc:title xml:lang="eng">Test DC eng title</dc:title>
                <dc:title xml:lang="eng">Test DC eng title 2</dc:title>
                <dc:creator>Test DC creator</dc:creator>
            </METS:xmlData>
        </METS:mdWrap>
    </METS:dmdSec>
    <METS:dmdSec ID="description">
        <METS:mdWrap MDTYPE="MODS">
            <METS:xmlData>
                <mods:mods xmlns:mods="http://www.loc.gov/mods/v3" xsi:schemaLocation="http:
                    <!-- Pri titleInfo je atribut @type opcijski,
                         nujen pa pri identifikaciji alternativnih naslovov (alternativeTitl
                         Kot alternativni naslov sta možna translated in alternative
                    -->
                    <mods:titleInfo>
                        <mods:title xml:lang="slv">Testna MODS publikacija</mods:title>
                        <mods:title xml:lang="eng">Test MODS publication</mods:title>
                        <mods:subTitle xml:lang="slv">Podnaslov</mods:subTitle>
                        <mods:title xml:lang="eng">Subtitle</mods:title>
                        <mods:title xml:lang="deu">Title (deutsch)</mods:title>
                    </mods:titleInfo>
                    <mods:titleInfo type="alternative">
                        <mods:title xml:lang="eng">Alternative title of the publication in E
                    </mods:titleInfo>
                    <mods:titleInfo type="translated">
                        <mods:title xml:lang="deu">Übersetzung des Titels in deutscher Sprac
                    </mods:titleInfo>
                    <!-- Primer za avtorja -->
                    <!-- Za imena je priporočljivo označiti atribut type -->
                    <mods:name type="personal">
                        <!-- Osebno ime se vedno zapiše v namePart kot Priimek, Ime  -->
                        <mods:namePart>Priimek, Ime</mods:namePart>
                        <!-- Če elementu namePart dodamo @type, ti podatki ne bodo indeksira
                              npr. family, given, date, termsOfAddress -->
                        <mods:namePart type="family">Priimek</mods:namePart>
                        <mods:namePart type="given">Ime</mods:namePart>
                        <mods:namePart type="date">Datum rojstva in smrti</mods:namePart>
                        <mods:namePart type="termsOfAddress">Naslov osebe</mods:namePart>
```

Xml | Result Si4 | XPath Evaluate Test

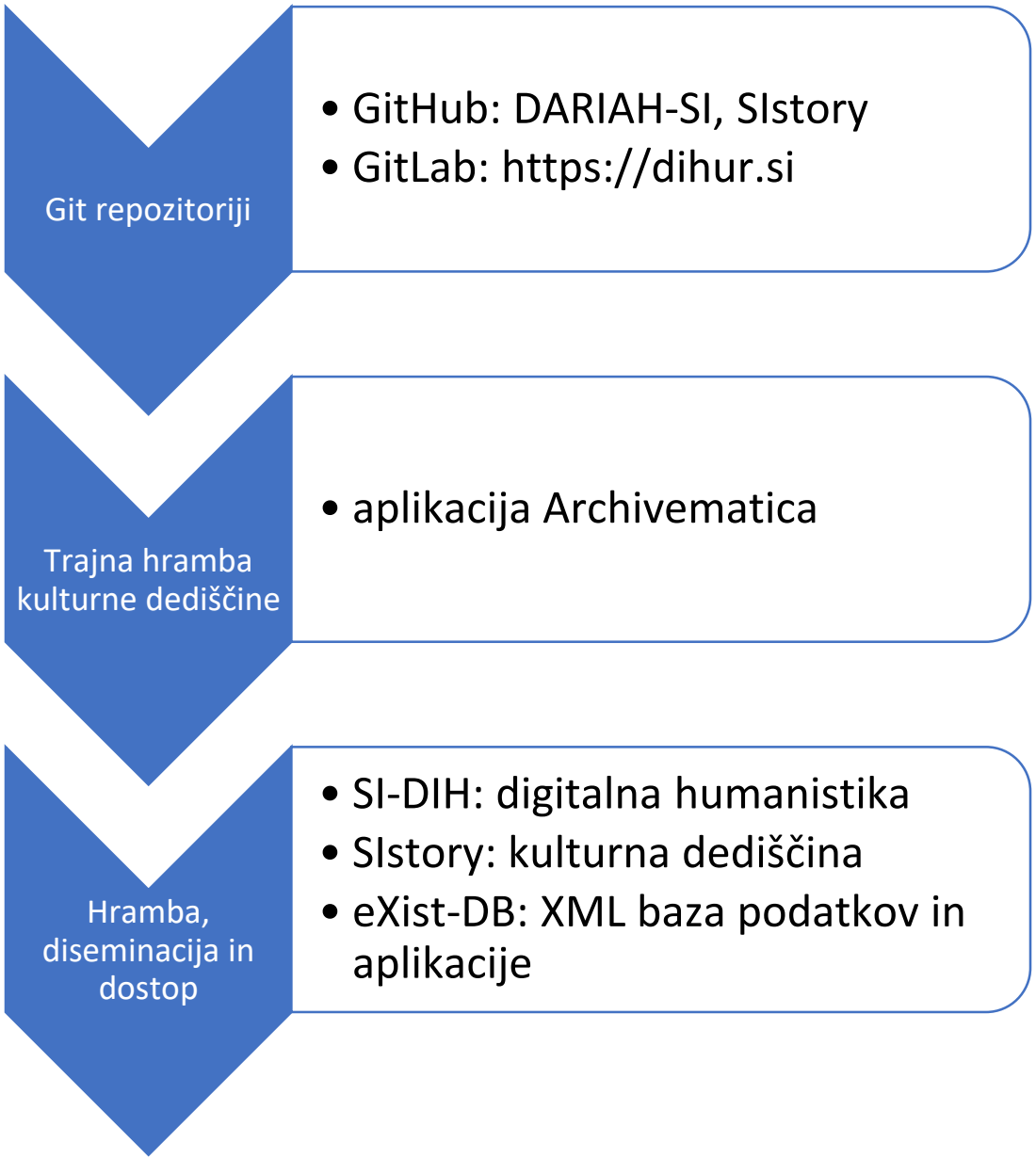

```json
{
  "header": {
    "id": "si4.test123",
    "objId": "http://hdl.handle.net/20.500.12325/test123",
    "type": "entity",
    "createDate": "2019-11-12T14:03:10",
    "lastModDate": "2019-11-12T14:03:10",
    "recordStatus": "Active",
    "creators": [
      {
        "name": "Pančur, Andrej",
        "type": "INDIVIDUAL",
        "id": "3"
      }
    ],
    "disseminators": [
      {
        "name": "Sidih",
        "note": "https://sidih.si",
        "type": "ORGANIZATION"
      }
    ]
  },
  "si4tech": {
    "additionalMetadata": "false"
  },
  "files": [],
  "si4": {
    "title": [
      {
        "metadataSrc": "dc",
        "value": "Test DC slv title",
        "lang": "slv"
      },
      {
        "metadataSrc": "dc",
        "value": "Test DC slv title 2",
        "lang": "slv"
      },
      {
        "metadataSrc": "dc",
        "value": "Test DC eng title",
        "lang": "eng"
      },
      {
        "metadataSrc": "dc",
        "value": "Test DC eng title 2",
        "lang": "eng"
      },
      {
        "metadataSrc": "mods",
        "value": "Testna MODS publikacija",
        "lang": "slv"
      },
      {
        "metadataSrc": "mods",
        "value": "Test MODS publication",
        "lang": "eng"
      },
      {
        "metadataSrc": "mods",
        "value": "Subtitle",
        "lang": "eng"
      },
```

Mapiranje iz ElasticSearch JSON v sprednji del spletne strani, polja za napredno iskanje in OAI-PMH (spodaj primer)



Infrastrukturna podpora za raziskovalne podatke digitalne humanistiki v Sloveniji:

- Projekti večinoma potekajo v Git repozitorijih za kontrolo verzij.
- Digitalne izdaje kot statične HTML spletene strani z dinamično vsebino => optimalno za trajno hrambo in vzdrževanje.
- Digitalizirano kulturno dediščino trajno hranimo v aplikaciji Archivematica (ISO-OAIS fukcionalni model). Ni javno dostopna.
- Javno dostopni rezultati projektov in raziskovalni podatki v repozitoriju SI-DIH, SIstory, različnih bazah podatkov.



Načrti in želje

- Pisanje dokumentacije in certificiranje repozitorija za raziskovalne podatke za digitalno humanistiko SI-DIH
- Aplikacija za predajo raziskovalnih podatkov: datotek in metapodatkov
- Git kontrola verzij METS XML datotek
- SWORD za izmenjavo dostavnih informacijskih paketov med Archivematico in si4
- Mapiranje si4 v LOD (Linked open data) in priprave za prehod na semantični splet
- Dodatne si4 entitete: akterji, dogodki ipd.