DOKUMENTOINTI JA METADATA

Perusteita, problematiikkaa, hyviä ja huonoja esimerkkejä

Mari Elisa ”MEK” Kuusniemi,
ORCID: 0000-0002-7675-287X
Helsingin yliopiston kirjasto
Tutkimuksen palvelut

Mitä kaikkea on tutkimusdatan metadata?

Mitä tarkoittaa dokumentaatio?

Kuvaillaanko tutkimusdataa?
Onko siihen olemassa työkalua?

Onko dokumentaatio eri asia kuin metadata?

Mitä tarkoittaa dokumentointi?

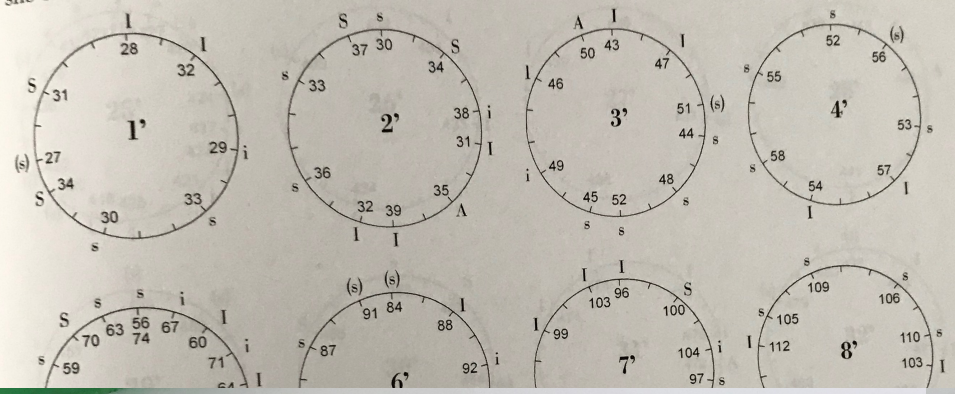
Mikä ihme “*data provenance*” on? Liittyykö se asiaan?

Mitä metadatastandardi tarkoittaa tutkimusdatan kontekstissa?

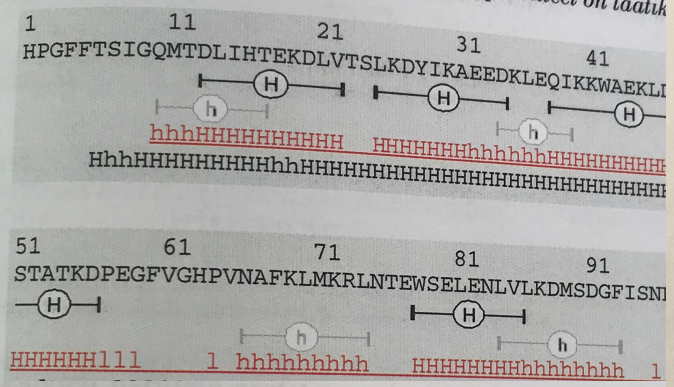
Opetetaanko jossain tutkimusdatan dokumentointia?

Liite 4. α-kierteiden toteaminen Edmundson-renkailla.

Kehien sisäpuolella olevat pienet numerot tarkoittavat aminohappojärjestystyksen rinnakkainasettelun kohtia. Kehän ulkopuolella olevat kirjaimet tarkoittavat aminohappotähteiden ennustettuja sijainteja. I/i = proteiinin sisällä, S/s = proteiinin pinnassa, A/a = aktiivisessa kohdassa. Pienet kirjaimet kuvaavat epävarmempia ennusteita. Sulkeissa olevat sijainnit on ennustettu "käsin", siis ei Darwin-järjestelmällä.



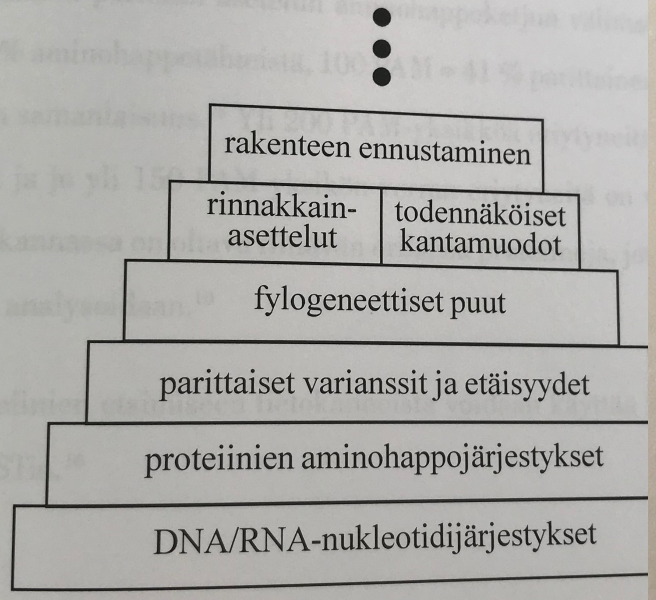
Kuva 6. Ihmisen prolyyli-4-hydroksylaasin α(I)-alayksikön ennustettu sekund
naalipeptidi on poistettu. H (helix) = α-kierre; E (extended) = β-juoste; pienet
nä epävarmaa ennustetta. ETH-menetelmällä laadittu ennuste on esitetty pu
rirakenteen elementtien alku- ja loppukohdat on merkitty viivoilla, epävarmat
na. Riveillä alimpina (musta teksti) on Heidelbergin järjestelmällä tuotettu ve
hkaisemmissa tutkimuksissa aktiivisiksi todetut aminohappotähteet on laatik



Liite 3. Heidelbergin täysautomaattisen järjestelmän tuottama prolyyli-4-hydroksylaasin α-alayksikön sekundäärirakenteen ennuste.

Tässä on esitetty sellaisinaan Heidelbergin järjestelmän tulosteesta sen valitsemien aminohap järjestysten esittely sekä varsinainen ennuste.

```
--- ----------------------------------------------------------
--- MAXHOM multiple sequence alignment
--- ----------------------------------------------------------
---
--- MAXHOM ALIGNMENT HEADER: ABBREVIATIONS FOR SUMMARY
--- ID            : identifier of aligned (homologous) protein
--- STRID         : PDB identifier (only for known structures)
--- PIDE          : percentage of pairwise sequence identity
--- WSIM          : percentage of weighted similarity
--- LALI          : number of residues aligned
--- NGAP          : number of insertions and deletions (indels)
--- LGAP          : number of residues in all indels
--- LSEQ2         : length of aligned sequence
--- ACCNUM        : SwissProt accession number
--- NAME          : one-line description of aligned protein
---
--- MAXHOM ALIGNMENT HEADER: SUMMARY
ID          STRID  IDE WSIM LALI NGAP LGAP LEN2 ACCNUM NAME
p4ha_human          100  100  534    0    0  534 P13674 PROLYL 4-HYDROXYLASE ALPH
p4h1_mouse           94   96  526    6    6  534 P47867 PROLYL 4-HYDROXYLASE ALPH
p4ha_rat             96   98  534    0    0  534 P09245 PROLYL 4-HYDROXYLASE ALPH
p4ha_chick           87   92  513    3    7  534 P54001 PROLYL 4-HYDROXYLASE ALPH
p4h2_mouse           62   78  520    8   14  516 P16924 PROLYL 4-HYDROXYLASE ALPH
```



Kemistien Internetin kautta tapah tiedonhankinnan ja viestinnän kar Oulun yliopiston kemian laitoksell

Kartoitamme Oulun yliopiston kemian laitoksella kemistien ja erikoisty opiskelijoiden Internetin kautta tapahtuvaa tiedonhankintaa ja viestintää informaatiotutkimuksen opinnäytetyötä.

Pyydämme vastaamaan seuraavassa esitettyihin kysymyksiin. Enimmäk napautus.

Lomakkeen lähettämisestä on ohje lopussa. Vastaukset käsitellään luotta

VASTAAJAN KUVAUS:

1. Sähköpostiosoite:

2. Sukupuoli:
○ mies ○ nainen

3. Syntymävuosi:

4. Vastaaja työskentelee:
- ○ epäorgaanisen kemian osastolla
- ○ fysikaalisen kemian osastolla
- ○ orgaanisen kemian osastolla
- ○ rakennetutkimuksen osastolla

5. Vastaajan työskentelee kemian laitoksella:

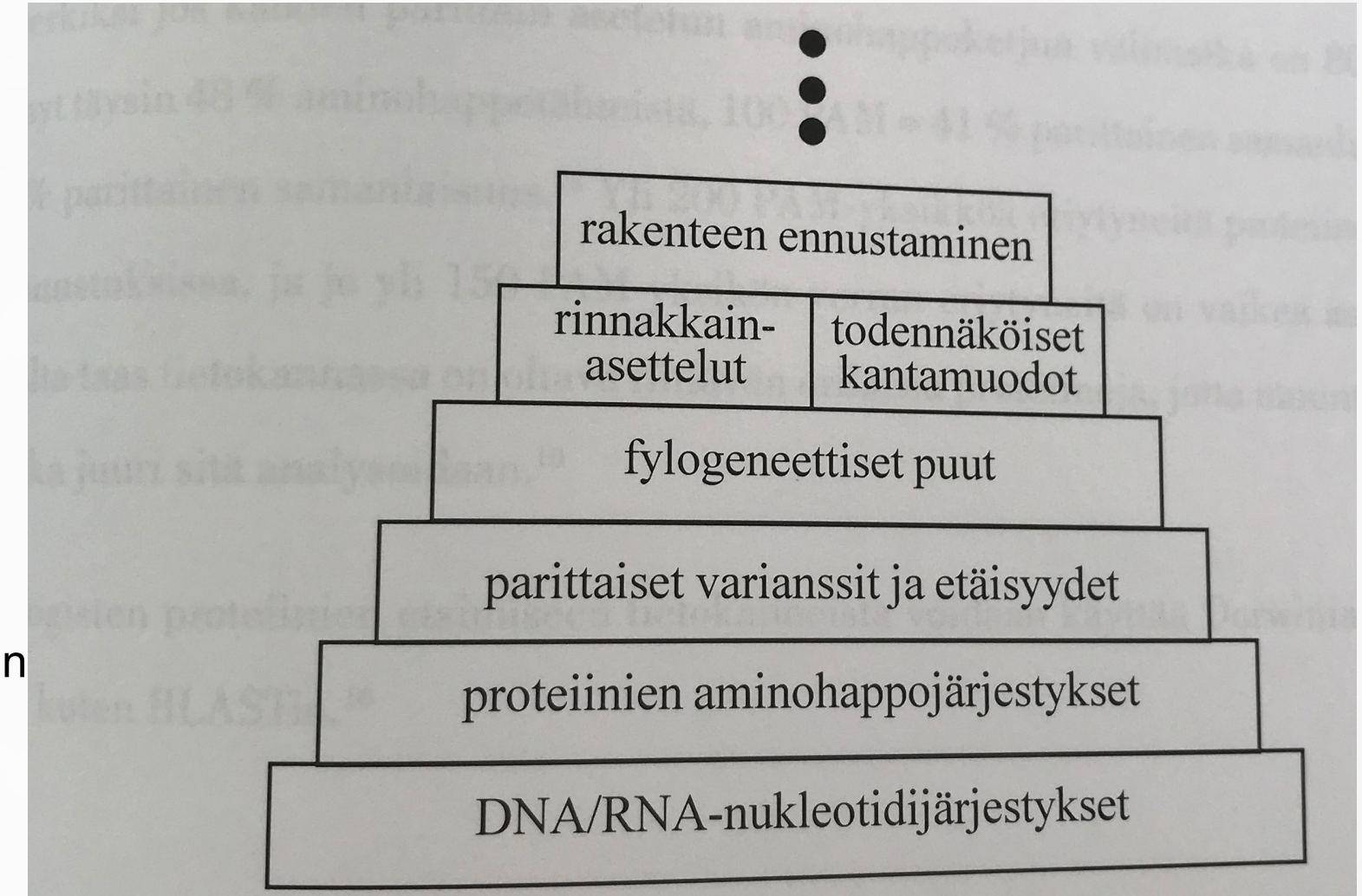
- ☐ amanuenssina
- ☐ apulaisprofessorina
- ☐ assistenttina
- ☐ dosenttina
- ☐ erikoistyöntekijänä
- ☐ laboratorioinsinöörinä
- ☐ opintoneuvojana
- ☐ professorina
- ☐ päätoimisena tuntiopettajana
- ☐ tutkijana
- ☐ yliassistenttina
- ☐ yli-insinöörinä

minä?

Lähde: MEKin kemian (bioinfomatiikkaa) sekä informaatiotutkimuksen gradut vuodelta 1999

Yhden kemian gradussa käytetyn menetelmän informaatiopyramidi

- Kaikki tasot käyttävät ja tuottavat dataa.
- Jokainen taso vaatii oman erilaisen metadatansa.
- Jokainen taso on kuvailtava, jos haluaa ymmärtää mistä tulos syntyi.



Darwin-järjestelmän informaatiopyramidi.[13]

MEKIN MIELIKUVA

Tutkimusdata
Menetelmät
Kuvailun kohteita
Koodit
Laskukaavat
Provenance

Skeemat
Standardit
Formaatit
Kuvailuohjeita
Tietomallit
Nimeämiskäytännöt

Dokumentointi
Kuvailu
Kuvailun tekemistä
Koodi- ja päiväkirjojen kirjoittaminen
Julkaisun yhteydessä yleiskuvaus

Dokumentaatio
Datan koodikirja
Kuvailutuotteita = metadata
Datasanakirja
Datapäiväkirja
Readme

METADATA

Metadata is "data that provides information about other data".[1] In other words, it is "data about data".

Many distinct types of metadata exist, including

- descriptive metadata,
- structural metadata,
- administrative metadata,[2]
- reference metadata,
- statistical metadata[3] and
- legal metadata.

Source: https://en.wikipedia.org/wiki/Metadata


Documentation is the practice where you label your data. Metadata is the text in the labels.
Foto by: @JoBrodie

Dokumentoinniksi kutsutaan sitä käytäntöä, jossa laputat purnukat.
Metadata on sitä mitä lapuissa lukee.

Foto by: @JoBrodie

METADATASTA PUHUMINEN EI OLE HELPPOA

- Dokumentointiin ja metadataan liittyvät mielikuvat ja käytetyt sanat riippuvat tutkijan kokemuksista ja ovat siksi aika yksilöllisiä.
- Keskustelutilanteessa on vaarana, että puhutaan samoilla sanoilla, mutta tarkoitetaan eri asiaa.
- Varsinaista dokumentoinnin koulutusta ei juuri ole tarjolla. Yleensä se opitaan kaiken muun ohessa.
- Hyvää dokumentointi vaaditaan kaikkialla, mutta kuka osaisi neuvoa siinä?
- Teimme vuonna 2018 tutkimusdatan dokumentoinnista oppaan, jotta asiasta keskusteleminen helpottuisi.

Tutkimusdatan metadatan perusopas

Käytetty mm. Fairdata PAS -palveluun säilytettäväksi siirrettävien data-aineistojen arviointilkeskusteluissa.

Luo yhteisen pohjan keskustelulle tutkimusdatan metadatasta.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Making a research project understandable

Guide for data documentation

Siiri Fuchs & Mari Elisa Kuusniemi
Helsinki University Library, Data Support

Thanks to Liisa Siipilehto (Helsinki University Library, UH), Juuso Ala-Kyyny (UH), Katja Moilanen (Finnish Social Science Data Archive, FSD), Mari Kleemola (FSD), Jessica Parland-von Essen (CSC), and Arto Teräs (CSC) for valuable feedback, when making this guide.

Version 1.1, 27.11.2018
DOI: 10.5281/zenodo.1684941

Oppaassa on kuvattu yleisesti käytettyjä tutkimusdatan metadatan peruselementtejä.

Tämä on vain yhdenlainen esimerkki siitä miten peruselimentit voidaan jaotella.

Opas on suunnattu kaikkien alojen tutkijoille, siksi se on tieteenalariippumaton.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

ELEMENTS OF DATA DOCUMENTATION

Metadata standards & vocabularies
- Describe data in a controlled format using vocabularies
- Use field and disciplinary specific standards, if suitable standards exist
- Should always be favored, if suitable for project

Data management software, i.e. databases and electronic laboratory notebooks
- Software take data in and create a database
- Makes documentation easier as software usually generate metadata by themselves
- Easy to share & control access, usually have safe storage & search tools
- Easy error spotting: inputs out of range can be automatically detected

Data dictionaries
- Dictionaries explain variables used in a dataset
- Codebooks are collections of codes, algorithms and calculations used in a project

Directory structure
- Create a folder structure to suit your project needs
- If you work with sensitive data, a clear folder system helps also in access control
- Balance between shallow and deep folder hierarchy to keep files findable

Tagging files
- Tags are keywords assigned to files, which enable organizing and searching files easier
- A file can only be in one folder at a time, but it may have an unlimited number of tags

File naming conventions
- Create a meaningful but brief system with unique names (in case of directory structure break-down) in the beginning of the project

Version control
- Version control makes it possible to return to an older version of a specific file
- Automatic version control system preferred

Readme-files
- Readme-files are text documents (e.g. format.txt) providing information about data files to ensure they are interpreted correctly
- Readme-files can include information such as title, creator, description, location, methodology, dates and file formats

Discovery metadata
- Descriptive metadata, "label of the dataset", where the dataset is explained should always be published regardless of the nature of the data
- Persistent identifiers (PIDs) identify citable online resources providing a permanent link to them. Persistent identifiers are used when citing and managing data and information

Research records
- Administrative documents (i.e. licenses, usage agreements, consents used) and other research related documents (e.g. research plan, publications, DMP) explaining the context of the actual metadata

ELEMENTS OF DATA DOCUMENTATION

Metadata standards & vocabularies

- Describe data in a controlled format using vocabularies
- Use field and disciplinary specific standards, if suitable standards exist
- Should always be favored, if suitable for project

Data management software, i.e. databases and electronic laboratory notebooks

- Software take data in and create a database
- Makes documentation easier as software usually generate metadata by themselves
- Easy to share & control access, usually have safe storage & search tools
- Easy error spotting: inputs out of range can be automatically detected

Data dictionaries

- Dictionaries explain variables used in a dataset
- Codebooks are collections of codes, algorithms and calculations used in a project

Directory structure

- Create a folder structure to suit your project needs
- If you work with sensitive data, a clear folder system helps also in access control
- Balance between shallow and deep folder hierarchy to keep files findable

Tagging files

- Tags are keywords assigned to files, which enable organizing and searching files easier
- A file can only be in one folder at a time, but it may have an unlimited number of tags

File naming conventions

- Create a meaningful but brief system with unique names (in case of directory structure break-down) in the beginning of the project

Version control

- Version control makes it possible to return to an older version of a specific file
- Automatic version control system preferred

Readme-files

- Readme-files are text documents (e.g. format.txt) providing information about data files to ensure they are interpreted correctly
- Readme-files can include information such as title, creator, description, location, methodology, dates and file formats

Discovery metadata

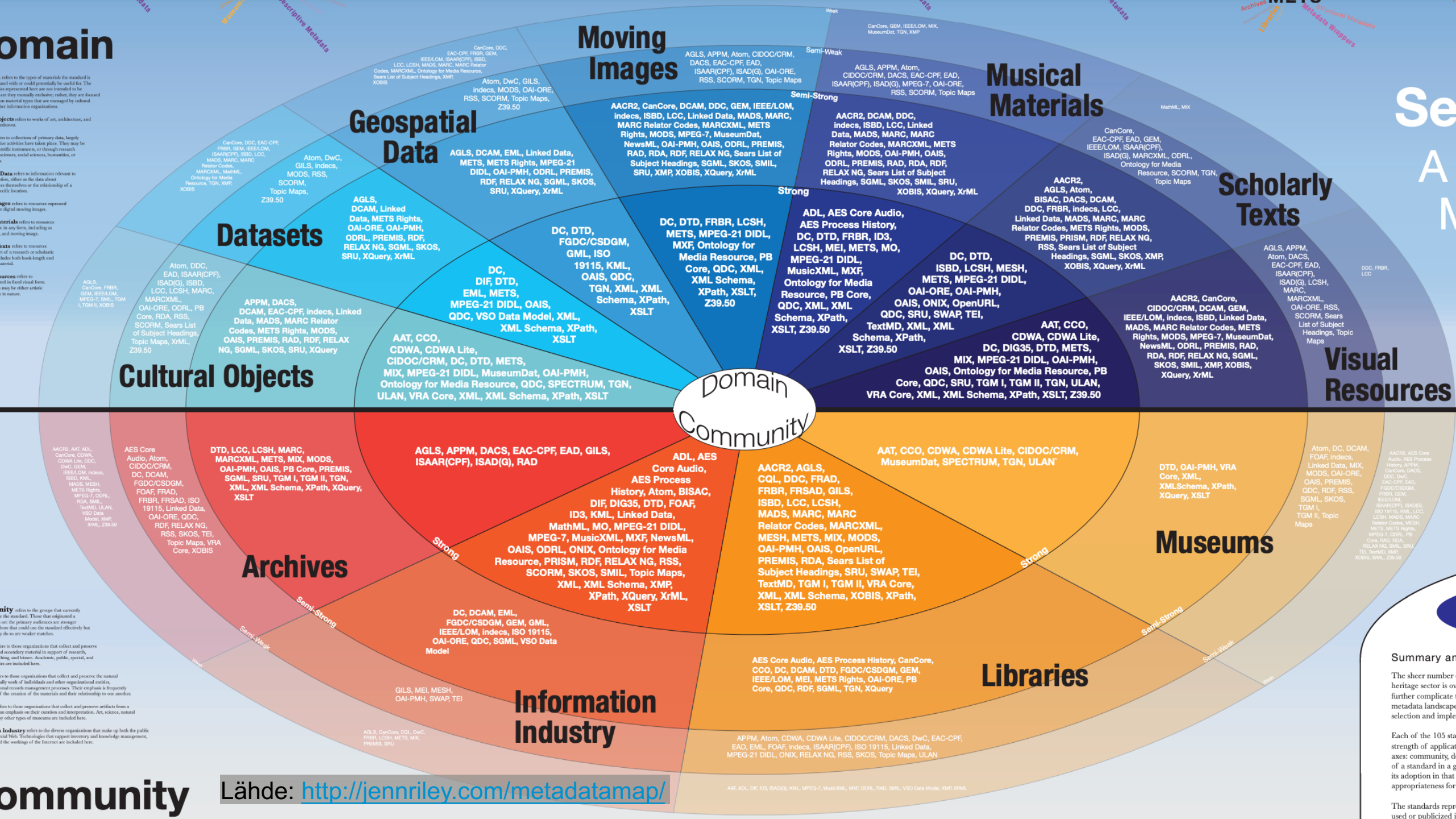
- Descriptive metadata, "label of the dataset", where the dataset is explained should always be published regardless of the nature of the data
- Persistent identifiers (PIDs) identify citable online resources providing a permanent link to them. Persistent identifiers are used when citing and managing data and information

Research records

- Administrative documents (i.e. licenses, usage agreements, consents used) and other research related documents (e.g. research plan, publications, DMP) explaining the context of the actual metadata

MITEN VOISI PAREMMIN TUTUSTUA TUTKIMUSDATAN METADATAAN?

- Etsi joitain tutkimusdatan metadatastandardeja ja ihmettele niitä.
- Standardeja on helppo löytää (haku: datatyyppi + metadata standard).
- Standardeja on vaikea ymmärtää. => Ei ole mikään ihme, ettei niitä tutkijoiden arjessa useinkaan käytetä.
- Ymmärtäminen ei ole kuitenkaan mahdotonta, mutta sinnikkyyttä se vaatii.
- Standardi voi osoittautua helpoksi seurata, kuhan siihen ensin tutustuu.
  - Joskus on tarjolla työkaluja, jotka tukevat standardin käyttöä.



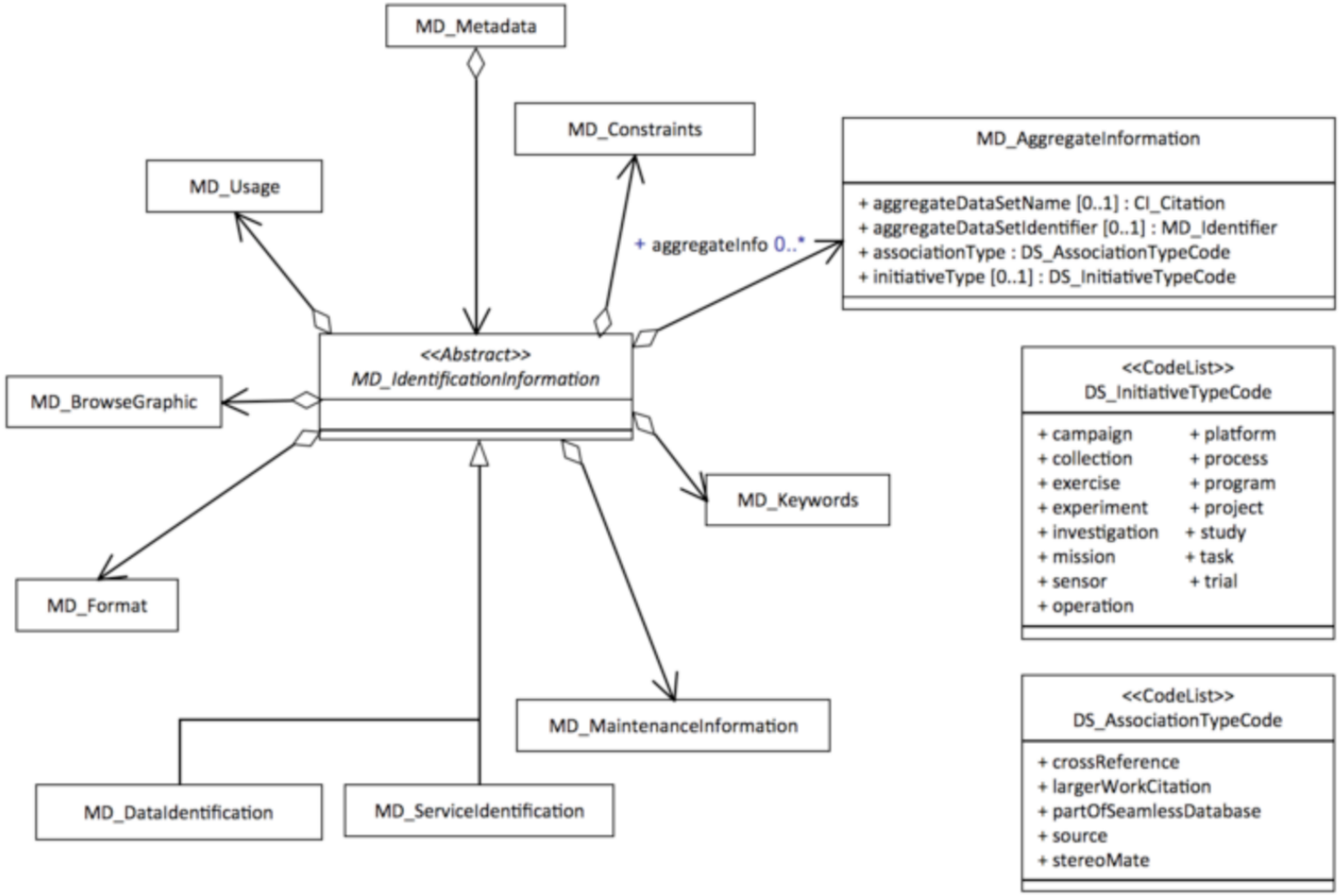
Lähde: http://jennriley.com/metadatamap/

Publications of the ISO Geospatial Metadata Standard

Technical and Formal Publications

ISO standards must be purchased. The American National Standards Institute (ANSI) serves as the US member agency to ISO and provides easier access to the standards and, generally, at a lower cost.

- ANSI eStandards Store - Search for an 'INCITS/ISO' version of the standard to get the best price, e.g. 'INCITS/ISO 19115-1' $133 (pdf) as of August 2015
- ISO Store



Supporting Publications and Representations

- ISO/TC 211 Harmonized Model Web Server - Access to UML diagrams for ISO/TC 211 standard
- NOAA-NCEI Metadata - ISO 19115:2003 Geographic Information - Metadata Workbook (2.5 MB) - Guide to Implementing ISO 19115:2003(E), the North American Profile (NAP), and ISO 19110 Feature Catalogue.

METADATA JA FAIR

- FAIR periaatteet tavoittelevat koneluettavuutta.
- FAIR on mielikuvissa muotoutunut tarkoittamaan myös ihmiselle ymmärrettävää dataa.
- FAIR evaluointityökalut (esim. F-UJI https://www.f-uji.net/) mittaavat koneluettavuutta.
  - EOSC-Nordic projektissa olemme kokeilleet kahta FAIR-evaluointityökalua oikeille data-aineistoille. Arviointityökalua antaa FAIR-pisteet datarepositoriolle.
  - FAIR-evaluointityökalut eivät mittaa sitä onko data käytettävää jatkotutkimuksessa (ymmärrettävää ihmiselle).
  - DOI on osoittautunut tehokkaaksi FAIR-pisteiden nostajaksi, koska DataCite tarjoaa minimetadatan DuplinCore-muodossa avoimen rajapinnan kautta (jota evaluointityökalut ymmärtävät).
  - FAIR näyttäytyy käytännössä tavoitteena, joka edistää datan jatkokäyttöä lähinnä datan löytyvyyden näkökulmasta.

METADATA JA FAIRDATA PAS

- Fairdata PAS-palvelun käyttö vaatii korkeakoululta datan arvonmääristystä.
- PAS-prosessi on hyvä mahdollisuus jutella tutkijoiden kanssa datan dokumentoinnista ja metadatasta.
- Pitkäaikaissaatavuus vaatii itsensä selittävän data-aineiston eli datan dokumentaatio pitäisi olla niin hyvä, ettei data keränneisiin tutkijoihin tarvitse olla yhteydessä sitä käyttääkseen.
  - ☞ Selkeä lähtökohta keskustelulle.
  - ☞ Usein huomataan, että
    - ☞ dokumentointikäytännöt pitäisi suunnitella paremmin,
    - ☞ jälkikäteen täydentäminen on työlästä ja rahoitusta sille työlle on hankala löytää.
  - ☞ Päätetään suunnitella ja dokumentoida paremmin seuraava hanke.

TUTKIMUSDATAN DOKUMENTOINNISTA PUHUMINEN EI AINA OLE HELPPOA, MUTTA HAUSKAA JA MIELENKIINTOISTA SE ON!

“Coming Out of Your Comfort Zone: A Tough Decision”

“…it’s not the specific role or job title that librarians have that will make a difference, but their attitude.”

“…in order to be heard, you have to go where the researchers are and talk to them in a language that they understand.”

- *Stein Høydalsvik*

Source: Papadopoulou, Elli. (2019) Chapter 8.2. Starting at the End: Seniors’ Research Data Project at the UiT The Arctic University of Norway, *Engaging Researchers with Data Management - The Cookbook.* DOI: 10.11647/OBP.0185

KIITOS!

mari.elisa.kuusniemi@helsinki.fi