SND
Svensk nationell datatjänst

Mer om metadatastandarder

Pass 4: Metadatastandarder

BAS Online 2021-01-20

Välkommen till presentation 3 i pass 4. Den här presentationen handlar om några olika teman som har att göra med metadata-standarder.

Jag kommer att börja prata lite om hur man kan mappa mellan olika standarder så att de blir kompatibla med varandra. Sen kommer vi avhandla två sätt att precisera sitt användande av metadatastandarder, nämligen hur man kan skapa egna profiler från en metadatastandard och hur man använder kontrollerade vokabulärer för att standardisera informationen i en metadata-beskrivning.

Forskare använder sällan metadatastandarder själva och är ofta inte intresserade av att sätta sig in i en komplicerad metadata-standard med oklar nytta för dem. De som behöver kunna meta-datastandarder är i första hand vi som jobbar med katalogisering av forskningsdata. Vi behöver ha grundläggande kunskaper i hur man använder metadatastandarder och vara medvetna både om fördelar men också om eventuella svårigheter med att använda standarder. Att kunna alla standarder är, som vi var inne på i förra presentationen, en omöjlighet och eftersom det finns så otroligt många standarder att välja mellan är det inte alldeles enkelt att veta var man ska börja leta.

När man ska välja en metadatastandard att jobba med finns det en del saker att ta hänsyn till. Olika discipliner har olika behov när det gäller metadata och därför finns det också många olika metadata-standarder och olika sätt att använda dem.

Vilken typ av dokumentation behöver göras för att data ska gå att förstå och återanvända? Den här frågan kan man ofta bara besvara efter att ha pratat med en forskare med ämneskompetens som kan fundera över vad den själv skulle behöva för att kunna återanvända en kollegas fil.

Inte alla metadatastandarder stöder flerspråkiga metadata. Det är viktigt att känna till om man vill ha både nationell och internationell synlighet för de data man beskriver.

Sist men inte minst är det viktigt att fundera på om det redan finns färdiga verktyg för att jobba med standarden och om verktyg som används av forskaren har stöd för någon viss metadatastandard. Verktyg är viktiga eftersom det är väldigt tidskrävande att arbeta med en metadatastandard utan ett passande verktyg. Och om forskaren kan exportera metadata från sitt verktyg i ett visst format kan det spara mycket arbete jämfört med att behöva mata in alla metadata själv.

Digital Curation Centre[1] och Research Data Alliance[2] har listor över metadatastandarder för forskningsdata och mer information.

Mappning mellan standarder

Eftersom olika discipliner kräver helt olika typer av metadata är det sannolikt att man måste arbeta med flera metadatastandarder. För att det ska gå så smärtfritt som möjligt bör det finnas en mappning mellan de standarder man använder. Det innebär att man slår fast vilket element i den ena standarden som motsvarar vilket element i den andra. När man har gjort det kan man ofta ganska enkelt importera och exportera metadata mellan olika system. På det sättet kan en sökning på ett ställe ge resultat från många olika ämnesområden.

Att göra en mappning själv ger också bättre förståelse för en standards möjligheter och begränsningar. Det kan hjälpa en att välja

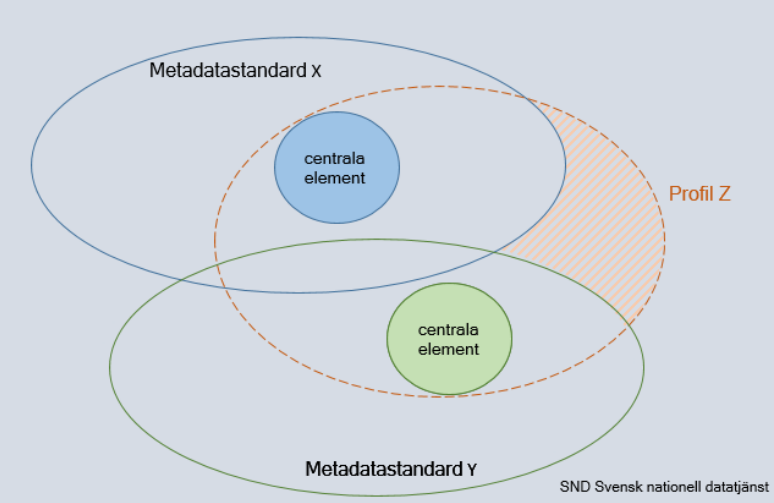
mellan flera standarder, eftersom det blir tydligt vilken standard som är mer detaljerad inom ett visst område.

När man ska mappa mellan två standarder jämför man de element som finns i varje standard med varandra. Man läser deras definitioner och konstaterar om de motsvarar varandra helt eller delvis. Samtidigt kan man också lägga märke till element som inte har någon motsvarighet alls.

Om man konstaterar att två element motsvarar varandra delvis så finns det några olika termer som närmare bestämmer på vilket sätt de motsvarar varandra. De här termerna visar att standarder inte alltid delar in information i samma kategorier:

- Exact match
- Close match
- Broad match
- Narrow match

De flesta, om inte alla, metadatastandarder tillåter ett visst mått av frihet när man ska tillämpa dem. Man kan välja bort vissa element eller bestämma sig för att tillämpa dem på ett snävare sätt än standarden föreskriver. I nästan alla fall måste man ändå ta hänsyn till standardens centrala element, de element som är standardens byggstenar och som resterande element i standarden bygger vidare på. Genom att välja ut element skapar man en metadataprofil för sin egen användning av standarden. Ibland kan det också vara så att man behöver kombinera flera metadatastandarder för att kunna uppfylla olika krav på sina metadata. Förutom de element som ingår i den eller de metadatastandarder som man bygger sin profil på, kan ytterligare element behövas för att beskriva materialet på bästa sätt.



Det är de som är markerade i den här bilden. För det mesta handlar det om administrativa metadata, som t.ex. tillgänglighet för data enligt den egna organisationens riktlinjer. Det kan också handla om ämnesspecifika krav där standarder ännu inte är särskilt omfattande.

För att skapa en metadataprofil behöver man veta vad man ska ha sina metadata till, med andra ord behöver man ta fram exempeldata med tillhörande dokumentation som man kan testa sin profil på. Man behöver ta hänsyn till om standarden har obligatoriska fält och man kan också behöva bestämma sig för om vissa fält som är frivilliga i standarden ska vara obligatoriska i ens eget arbete. På det här sättet skapar man en egen uppsättning regler för hur man vill följa standarden.

För att säkerställa att ens profil följs och att den inte är i konflikt med standarden kan man använda ett system för teknisk validering som automatiskt jämför den egna profilen och de egna metadatabeskrivningarna med standarden och varnar om det visar sig att det finns något fel. Om standarden till exempel föreskriver att man ska använda språkkoder enligt ISO 639-3 men man har fyllt i namnet på

språket istället kommer systemet för validering att varna så att man kan rätta till felet innan man publicerar metadata.

På SND har vi valt ut ett antal av DDI:s över tusen element och satt ihop våra profiler utifrån olika ämnesområden. Vi har också kompletterat med egna element som fyller funktioner som inte stöds av DDI. Våra profiler har tagit olika lång tid att utveckla och förändras fortfarande när nya behov uppstår. Eftersom SND är med och utvecklar DDI kan vissa förändringar som vi gör också integreras i nya versioner av DDI. I och med att det hos SND ska gå att beskriva data från alla olika ämnesområden ser vi över vilka andra standarder som är viktiga för våra ämnesspecifika profiler.

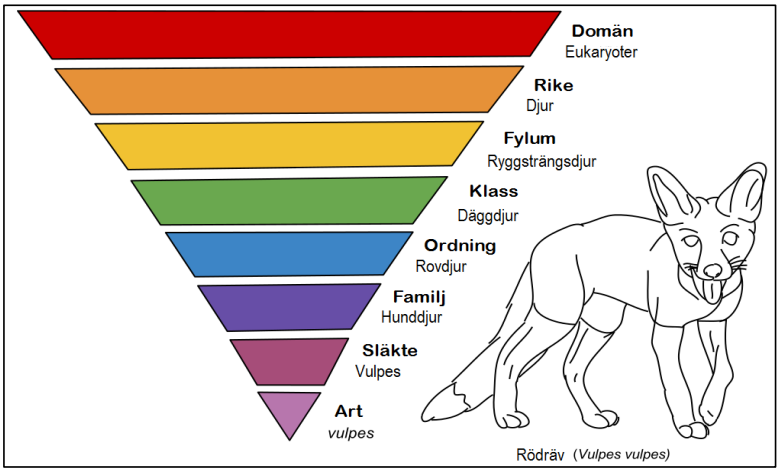
Kontrollerade vokabulärer

Kontrollerade vokabulärer är ett sätt att begränsa de möjliga värdena i ett element, man talar om vilken information som är tillåten för elementet. De består av listor med tillåtna värden och sedan kan man med hjälp av ett system för validering se till att endast de tillåtna värdena anges. Genom att använda kontrollerade vokabulärer underlättar man harmonisering mellan standarder och minskar risken för att fel värden fylls i.

Kontrollerade vokabulärer kan vara enkla listor. Här nedan är ett exempel på DDI-elementet *Kind of data format*. Metadata för det här elementet måste vara något av de termer som listas, exempelvis ”text”, ”stillbild” eller ”ljud”. Men det kan också vara mer komplexa listor, till exempel hierarkiskt ordnade taxonomier eller tesaurusar.

Kind of Data Format		Describes the physical format(s) of the data documented in the logical product(s) of a study unit.
Code	**Term**	**Definition**
Numeric	Numeric	Data consisting largely of values expressed as digits from 0 to 9 and, optionally, signs for negative values, decimal points, or letters only when intended to represent numbers (for example, A-F or a-f in hexadecimal).
Text	Text	Data consisting largely of text, including letters, numbers, and special characters or symbols used in writing for punctuation, abbreviation, etc. For example, interview transcriptions, narratives or essays written by study participants, newspaper articles, etc.
StillImage	Still image	Static images, such as graphs, drawings, photographs, diagnostic/medical images like X-rays, etc.
Geospatial	Geospatial	Geospatial data are any type of data with spatial coordinates that allow them to be mapped to the Earth's surface. They can represent physical objects, discrete areas or continuous surfaces. Discrete geospatial data are usually represented using vector data consisting of points, lines and polygons, while continuous geospatial data are usually represented by raster data, consisting of a grid of cells that each has its own value. Any number of applications in a wide range of areas produce geospatial data, such as GIS, Remote Sensing equipment, GPS units, archaeological total stations, manual mapping and computer-aided design (CAD), in a number of formats, including images, vector, text, and tabular data. Vector-based geospatial data include tables listing archaeological sites along with their coordinates, text-based files (e.g. XML) containing coordinates and topology for historic road networks, voting figures for political parties by administrative area. Raster-based geospatial data include satellite images, aerial photographs, scanned maps, and digital maps of elevations, vegetation, land-use, sea surface temperatures, air pollution, soil-types, etc.
Audio	Audio	Recorded sound, including voice, music, etc.
Video	Video	Moving images. May include films, animation, digital recordings, visual output from simulations, recorded television programs, etc. May be mute or may include synchronized sound.
Software	Software	Computer program(s) in source code (human-readable) or compiled form.
InteractiveResource	Interactive resource	A resource requiring interaction from the user to be understood, executed, or experienced. For example, training modules, query/response portals, files that require action from the user, etc.
ThreeD	3D	Virtual three-dimensional representations of objects, architecture, places, etc.
Other	Other	Use when the kind of data format is known, but not found in the list.

Här nedanför är ett exempel på taxonomi för rödräv hämtat från Wikipedia[3]. Taxonomier och tesaurusar kan innehålla värden med komplexa relationer till varandra. Sådana komplexa relationer kan vara att två värden liknar varandra eller är varandras motsats eller att det ena förekommer i ett sammanhang och det andra i ett annat.



Taxonomier 2018-03-06

Kontrollerade vokabulärer kan vara öppna eller slutna. Öppna kontrollerade vokabulärer kan fyllas på med nya värden vid behov, men det kan bara göras av en viss organisation eller annat kontroll-organ. Slutna kontrollerade vokabulärer kan inte ändras, till exempel

för att det inte går att föreställa sig några andra värden i ett visst sammanhang.

Här är några exempel på kontrollerade vokabulärer. Använd en söktjänst för att hitta mer information om dem på nätet.

- Datum (ISO 8601)
- Språk (ISO 639-3)
- Land (ISO 3166-1)
- Län- och kommunkoder (SCB)
- MESH
- Library of Congress Subject Headings
- Standard för svensk indelning av forskningsämnen 2011 (SCB)

Sammanfattning

I den här presentationen har jag gått igenom hur man kan arbeta med mappning så att man till exempel kan arbeta med flera olika metadatastandarder och enklare kan överföra information mellan olika system.

Jag har pratat om metadataprofiler och hur man kan tänka när man ska skapa sin eller organisationens egen profil. Några tips är att testa profilen med exempeldata och att titta på flera olika standarder, kanske kan man plocka element från flera håll.

Och för att få ut så mycket som möjligt av en metadatastandard är det också bra att använda sig av kontrollerade vokabulärer. En del standarder har specificerat vilka vokabulärer som ska användas för specifika element.

Referenser

[1]Digital Curation Centre:
https://www.dcc.ac.uk/

[2]Research Data Alliance:
https://rd-alliance.org/

[3]Taxonomi. I *Wikipedia*.
https://sv.wikipedia.org/wiki/Taxonomi (Hämtad 2021-01-20)