Voorstel Masterproef Informatica

Titel: Ontwikkeling van een opslagplatform voor analytische data in een laboratorium voor cultureel erfgoed.

Bedrijf

Naam: Koninklijk Instituut voor het Kunstpatrimonium (KIK, Brussel)
Tel: 02 739 68 46
Externe promoter(en): Wim Fremout
mailadres(sen): wim.fremout@kikirpa.be
Andere begeleiders: Hans Opstaele, Edwin De Roock
mailadres(sen): hans.opstaele@kikirpa.be, edwin.deroock@kikirpa.be

- Is dit de 1e masterproef in het bedrijf in samenwerking met onze opleiding? Ja~~/Nee~~
- Is er in het bedrijf inhoudelijke en informatica-technische begeleiding ? Ja~~/Nee~~
- Kan de student in het tweede semester (februari-mei) 3 dagen per week in het bedrijf/onderzoekscentrum aanwezig zijn om te werken aan de masterproef? Ja~~/Nee~~
- Adres waar de student zal werken: KIK, Jubelpark 1, 1000 Brussel

Doelstelling van het project

Dit project stelt zich tot doel om een platform voor geautomatiseerde centrale archivering te ontwikkelen waarin analytische wetenschappelijke data en de bijhorende metadata (informatie over het staal, staalvoorbereiding, instrumentele parameters en annotaties bij de meting) samen opgeslagen worden. Dit zal leiden tot een harmonisatie, vereenvoudiging en efficiëntieverbetering van de huidige werkwijze. Het volle potentieel van een archiveringssysteem voor analytische data, ligt niet enkel in hergebruik en interoperabiliteit op lange termijn, maar nog meer in de ontsluiting van deze gegevens via BALaT (de institutionele databank van het KIK) en via diverse "*digital resources*" die gecreëerd (zullen) worden tijdens het EU-gesponsorde IPERION-CH project.

Bestaande situatie en probleemstelling

Erfgoedwetenschappen omvatten het onderzoeksveld van conservatie/restauratie, kunstgeschiedenis en wetenschap in functie van conservatie van cultureel erfgoed. De data die gecreëerd wordt in dit veld is extreem divers. Dit project zal een specifiek onderdeel van documentatie van cultureel erfgoed uitdiepen: het archiveren, organiseren, presenteren en delen van analytische wetenschappelijke data. Enorme hoeveelheden spectra, chromatogrammen, chemische mappings en beelden worden verspreid en ongeorganiseerd bewaard op persoonlijke computers en servers, vaak in propriëtaire formaten. Dergelijke werkwijze voor opslag van cruciale data houdt grote risico's in op middellange tot lange termijn op vlak van verlies van belangrijke gegevens. Vaak weet enkel de analist waar het geanalyseerde staal is genomen van een kunstwerk, waar een bepaald bestand is bewaard en met welke parameters de analytische data werden opgenomen (metadata). Propriëtaire bestandsformaten zijn meestal gebonden aan een specifiek analytisch meetinstrument: als analytische instrumenten en de bijhorende computers vervangen worden, is verlies aan gegevens en metadata niet uit te sluiten. Er is een dringende nood aan open, bij voorkeur *human-readable*, bestandsformaten voor langetermijnopslag en voor interoperabiliteit en het delen van data tussen onderzoekers die verschillende softwareplatforms gebruiken.

Technologieën die aan bod kunnen komen

PHP, python, HTML, CSS, javascript, JSON, SQL, noSQL (bv. MongoDB, ElasticSearch), SPARQL

Omschrijving van de opdracht ($\pm \frac{1}{2}$ pagina)

In samenspraak met de medewerkers van het laboratorium en data management van het KIK, en de Europese collega's in het IPERION-CH project zal een opslagplatform voor analytische data ontwikkeld worden, gebruik makend van openbron software die reeds in het KIK en daarbuiten ontwikkeld zijn. In het bijzonder zal het opslagplatform SpecLib uitbreiden. Dit is een PHP-gebaseerd, databankloos webplatform in ontwikkeling in het KIK om spectrale bibliotheken in verschillende open en propriëtaire bestandsformaten online te delen. Diverse soorten gegevens en metadata worden samen bewaard in een JSON-gebaseerd format en kunnen geraadpleegd worden met simpele zoekopties. Import- en exportfilters pakken het probleem aan van de niet-compatibele bestandsformaten die

nodig zijn voor de softwareprogramma's gebruikt door verschillende onderzoekers. SpecLib is in ontwikkeling; een preview kan geconsulteerd worden op https://speclib.kikirpa.be/preview/ (momenteel enkel https). De code is nog niet online beschikbaar, maar zal binnenkort als openbron gepubliceerd worden.

Taken:

- Een installeerbaar pakket maken op GitHub: Momenteel is SpecLib een eenmansproject op een individuele computer. Om samenwerking te bevorderen, zal het opslagplatform voor analytische data op GitHub gehost worden en omgevormd tot een gemakkelijk te installeren pakket. De installatieprocedure wordt gedocumenteerd (in het Engels).
- Opschalen van het opslagplatform: SpecLib was specifiek ontwikkeld om kleine datasets te delen, waardoor voor een databankloos design werd gekozen op basis van JSON bestanden. Een institutioneel opslagplatform moet efficiënt blijven in het archiveren, consulteren en delen van zeer grote hoeveelheden analytische gegevens, en toch flexibel blijven op vlak van dataformaten en metadataschema's. In deze taak worden de voor- en nadelen van databankloze JSON, SQL-opslag en noSQL geëvalueerd. De gekozen oplossing zal dan geïmplementeerd worden.
- Zoekmogelijkheden zullen geïmplementeerd worden, om onderzoekers de mogelijkheid te bieden de volledige opslagplatform eenvoudig te doorzoeken op specifieke zoektermen.
- Import- en exportfilters maken voor de vele open en propriëtaire bestandsformaten die gebruik worden door de wetenschappelijke softwarepakketten voor het opnemen en bewerken van analytische data. Deze kunnen van nul geschreven worden (indien de specificaties voorhanden), maar er bestaan ook reeds veel bestandsconvertors voor propriëtaire formaten, vaak openbron in diverse programmeertalen (C, java, python, matlab, R,...). Deze bestaande projecten kunnen eventueel geport en geïntegreerd als plug-in. Als proof-of-concept zouden twee analytische bestandsformaten worden uitgekozen; door een eenvoudig plug-in systeem kunnen dan in de toekomst snel nieuwe filters worden toegevoegd.
- Een application programming interface (API) wordt ontwikkeld om eenvoudige toegang tot de analytische data en metadata mogelijk te maken vanuit andere softwarepakketten. Analytische data is slechts één aspect van erfgoeddata; andere projecten beogen het ontwikkelen van gespecialiseerde platformen voor samenwerking (ConservationSpace) of (kunst)wetenschappelijk onderzoek (ResearchSpace en BALaT). Deze platformen zijn echter niet ontwikkeld om complexe analytische data te ondersteunen en de veilige opslag op lange termijn ervan te verzekeren. Een API zal helpen om analytische data op een correcte, flexibele en krachtige manier aan deze platformen aan te leveren.

Mogelijke uitbreidingen en opties

- Import- en exportfilters maken voor meerdere analytische bestandsformaten.
- Implementatie van een systeem voor toegangsrechten voor individuele bestanden of groepen van bestanden.

Nog graag een antwoord op volgende vragen:

1. Welke vaardigheden verwacht je van de student die dit voorstel uitwerkt?
   - Een correcte analyse maken van de vereiste functionaliteit
   - Systematisch en vlot kunnen programmeren in PHP, python en/of andere relevante programmeertalen
   - Ordelijk kunnen werken code becommentariëren
   - Voldoende kennis van Engels

2. Veronderstel dat je dit werk laat uitvoeren door een werknemer uit je bedrijf. Welk profiel zou die werknemer dan bij voorkeur hebben ?
   Master informatica, industriëel ingenieur informatica met grondige kennis in programmeren en databanken. Affiniteit met analytische data uit een wetenschappelijk labo is een voordeel.

3. Binnen welke tijd zou je van die werknemer de resultaten verwachten?