VERSLAG DATA WORKSHOP 'WATER IN DE STAD'

DOEL EN VERLOOP VAN DEZE WORKSHOP?

In deze workshop werd dieper ingegaan op de databeschikbaarheid. Door te werken naar standaarden en/of principes kan een vlottere uitwisseling van data mogelijk gemaakt worden. De workshop draaide volledig rond deze standaarden en principes. Welke bestaan er? Welke zijn er te verkiezen waar en wanneer en voor welk type data? Hoe worden data best bijgehouden en met welke systemen best uitgewisseld? Of bekeken vanuit de Open Dei referentiearchitectuur ging deze workshop over de interfaces van Community (delen en samenwerken), Content/Context (datastandaarden) en Computing (opslag en gebruik).

Deelnemers kregen vooraf een vragenlijst om in te vullen en een taak; de resultaten hiervan werden toegelicht tijdens de workshop en op basis van discussies in 3 break-outs en een plenaire terugkoppeling achteraf, werd hier verder op ingegaan. De neerslag hiervan is terug te vinden in dit verslag.

Tegelijkertijd werden ook twee andere gerelateerde initiatieven nader toegelicht bij het begin van de workshop:

- OSLO en meerbepaald het OSLO traject AIR & WATER; en
- Data broker van AIV.

Zie de powerpoint op de kennishub voor meer gedetailleerde informatie.

FEEDBACK VRAGENLIJSTEN EN WORKSHOP

Huidige opslag van data

Uit de bevraging blijkt dat de dataopslag hier en daar nog gebeurt in Excel sheets, deze zijn minder "deelbaar" of machine readable, maar er is bereidwilligheid om dit te veranderen ten voordele van data-uitwisseling en standaardisatie.

Datastandaarden

Verschillende standaarden worden naar voren geschoven: IMKL/INSPIRE, OGC WFS/WMS, OGC WaterML, GeoJSON, "iets in JSON-LD", maar ook dat de standaard varieert naargelang leverancier en dat dit geen goede uitgangssituatie is. De keuze voor slechts één standaard wordt ook vermeld.

Centrale vs decentrale opslag

Over het algemeen is de boodschap dat de belangrijkste en de meeste data best centraal (maar wel in standaard formaat) zouden staan, maar dat lokale instanties hiernaast graag nog decentraal data bijhouden. Daar staat dan tegenover dat er vanuit Vlaanderen gekeken wordt naar de coördinatie van uniforme, maar decentrale databrokers.

Op basis van het huiswerk werd geïnventariseerd over welk type data deelnemers beschikken, wat ze ermee zouden willen doen, in welk formaat ze staan, onder welke standaarden ze ontsloten zouden worden en waar die data dan te vinden zouden zijn. Deze gegevens werden samengebracht om een schema op te stellen voor de break-outs om de discussie rond deze vragen te voeden. Daarnaast werd ook geïnventariseerd over welke zaken er best afspraken gemaakt worden binnen VLOCA onder alle stakeholders (zie ppt slide 32).

BREAK-OUTS

Methode

De deelnemers werden verdeeld over 3 break-outs. Elke break-out kreeg een aanzet tot beslissingsboom voorgeschoteld op een MIRO-board als een eerste probeersel van een stappenplan om te komen tot de beste oplossing om data beschikbaar te maken. Deze beslissingsboom werd overlopen aan de hand van een aantal vragen over de gewenste datastandaarden, eventuele voorkeuren van dataopslagtypes en pub/sub type. Om ergens te starten werden op basis van het ingediende huiswerk van de deelnemers een genudgde classificatie gemaakt tussen 4 types data om de discussie te stimuleren: sensor data, GIS data, modelresultaten en staalname data.

Belangrijkste conclusies

Datatypes

Het is duidelijk dat nog meer types data gezien worden door de deelnemers en dit afhankelijk van hun uitgangspunt. De voorgestelde opsplitsing is een goed begin, maar niet altijd even relevant. Tevens ging het over semantiek. Een opdeling kan nuttig zijn, maar mag ons niet vastzetten. In vele gevallen zijn de issues dezelfde voor de verschillende types.

Het belangrijkste bijkomende onderscheid dat aangehaald werd is **real-time versus niet-real-time of historische data**. Op die basis zouden bv. al sensordata onderscheiden kunnen worden van de andere 3 voorgestelde types. Ook alle eerder contextdata zijn niet-real-time.

Een andere onderscheidingsfactor zou dan contextdata of **back-end data van overheden versus andere, door data-bedrijven en sensoren geleverde data** (ook wel 'industriële data' genoemd) kunnen zijn. En wat met data-enrichment?

Ontologische opdelingen zoals bij WaterML zijn ook mogelijk. Voor OSLO zijn de volgende 3 parameters primordiaal: de plaats (locatie, context,...); de observatie data; en de context van het device. Die 3 moeten samen blijven. Dit relateert ook aan het belang van unieke Ids dat elders vermeld werd.

Datastandaarden

Algemeen is er een vraag naar **standaardisatie**; de standaard zelf doet er op zich niet toe, want theoretisch kan je mappen tussen de verschillende datastandaarden, maar dat heeft dan natuurlijk zijn economische plaatje. De realiteit nu is dat er vanalles en nog wat gemaakt wordt en dat bovendien de leveranciers zelf nog een transformatie moeten doorgaan en daadwerkelijk de standaard moeten leveren.

Daar staat tegenover dat **flexibiliteit** nodig is, want standaarden evolueren vandaag trager dan technologische evoluties in sensoren en data-opslag. Tussen beide moet een evenwicht zijn en backward compatibility is daarin zeer belangrijk.

Externe versus interne keuken

Datastandaarden zijn een externe kwestie, data-opslag is een interne kwestie. Enkel de externe kwestie vereist standaardisatie.

Eigen data-opslag

Eigen data-opslag is bijvoorbeeld niet interessant voor satellietdata, waar het verplaatsen van de gigantische hoeveelheid data niet aan te raden is. Bovendien moet eigen data-opslag een duidelijk doel hebben zoals het trainen van modellen, archivering, calibratie etc.

De IT infra moet flexibel genoeg zijn om de **uitwisselbaarheid en opslag te ontkoppelen** (cf. REST API).

Wat men intern gebruikt, kan al sterk bepaald zijn door historische keuzes en de bestaande use case(s): uitwisseling, monitoring, presentatie, exploratie, onderzoek etc. Sensor data is ook meestal niet bedoeld voor 1 use case. Idealiter kan een database om met verschillende data-types. Hoe kan je flexibel omgaan met **multi-gebruik** met andere noden, bijvoorbeeld cross-domain. Smart data wordt belangrijk!

Databases waarbij time-series verwerkt worden verschillen van databases waarbij dit niet nodig is. Zeker indien deze data snel achter elkaar (~seconde) binnenstromen.

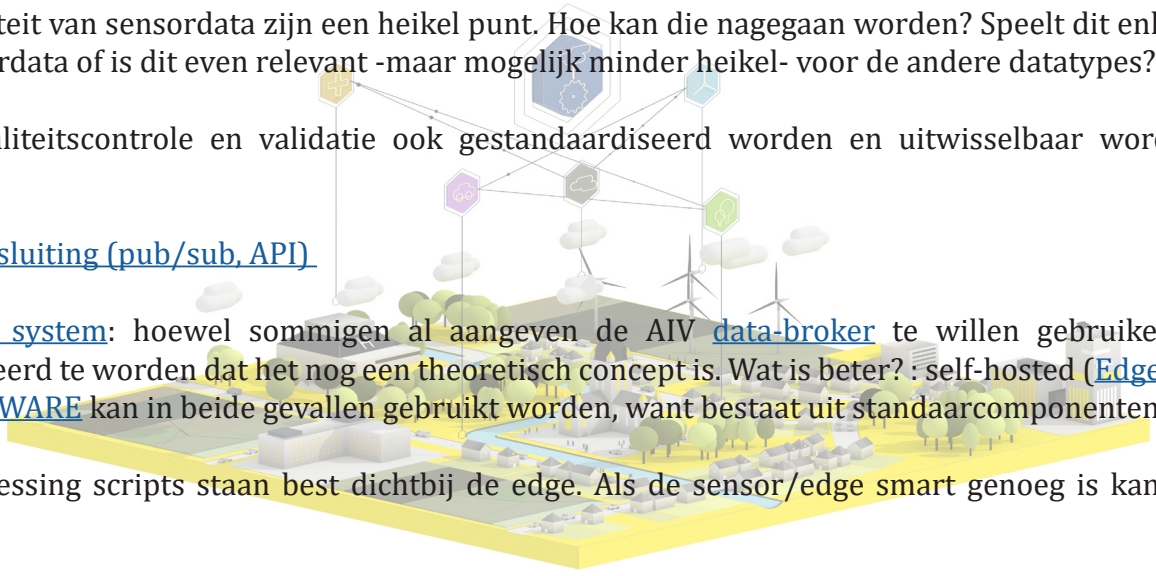
Zij die nog keuzes moeten maken, weten dan toch graag wat een goede keuze is voor data-opslag met een blik op de toekomst.

Quality control

De kwaliteit van sensordata zijn een heikel punt. Hoe kan die nagegaan worden? Speelt dit enkel maar bij sensordata of is dit even relevant -maar mogelijk minder heikel- voor de andere datatypes?

Kan kwaliteitscontrole en validatie ook gestandaardiseerd worden en uitwisselbaar worden (bv. scripts)?

Data-ontsluiting (pub/sub, API)



Pub/sub system: hoewel sommigen al aangeven de AIV data-broker te willen gebruiken, dient genuanceerd te worden dat het nog een theoretisch concept is. Wat is beter? : self-hosted (Edge) versus Cloud. FIWARE kan in beide gevallen gebruikt worden, want bestaat uit standaarcomponenten.

Pre-processing scripts staan best dichtbij de edge. Als de sensor/edge smart genoeg is kan daar al

voorverwerking gebeuren. Daar staat tegenover dat bepaalde klanten graag raw data hebben en andere reeds verwerkte. Kunnen beide opties in bepaalde gevallen open blijven?

Wat met message brokers/buffers (cf. Kafka)? In functie van de concrete toepassing, kan het nodig zijn dat informatie gebufferd wordt, omdat het verwerken ervan trager is dan de snelheid waarmee data binnen stroomt.

API: Er moet naar een interoperabel systeem van API gegaan worden (cf. OSLO verhaal). **Standaardisatie is hierin belangrijk en een directe vraag aan OSLO.** REST-API aub.

Versiebeheer is ook een belangrijk aandachtspunt.

Data-eigenschappen

Bestandgroottes blijken minder een issue te zijn.

Frequentie is zeer case-afhankelijk, maar heel hoge frequentie (meer dan 1/15min) is in het thematisch domein van water mogelijk minder relevant.

Gevoeligheid van data dient ook in het oog gehouden te worden (privacy, maar bv ook typisch voor water: puntlozingen en -vervuilingen die één-op-één terugtraceerbaar zijn).

Niet alle data hoeft zomaar **open** te zijn. Er is een verschil tussen open data en open-en-blote data (open by request).

Prijs van databeschikbaarheid

Ook het economische plaatje van al die databeschikbaarheid en met uitbreiding het hele IoT verhaal voor water en wie welk deel dan wel dient te betalen kwam in vele discussies bovendrijven.

In het kielzog van dit issue werd ook **linked data fragments** aangehaald.

Ten laatste moet er ook rekening gehouden worden met de noden van de vele andere watergebruikers die op dit moment niet vertegenwoordigd worden in dit traject. Het traject is zeer ruim, maar toch wordt er door de aanwezigen en hun projecten maar een deel van de noden die er zijn met betrekking tot de verschillende watertypes (grondwater, waterlichamen, drinkwater, afval water, hemelwater) afgedekt: waterverbruikende bedrijven (specifieke KMO's en industrie), landbouw (akkerbouw, veeteelt, tuinbouw), visserij, toerisme en recreatie, gezinnen en transport.

VOORUITBLIK

Volgende workshop zal verder ingegaan worden op het Cyber en Connection luik van de Architectuur: sensoren.