

Uitwisseling van brongegevens, Bron- en HDF5-format

Inleiding

Dit document bevat een beknopte toelichting op uitwisseling van brongegevens, via het Bron- en HDF5-format. Dit format en datamodel zijn op initiatief van de provincies en in opdracht van het programmabureau BRO ontwikkeld. Gebruik ervan staat echter vrij voor alle partijen, het is in die zin een open datamodel waar geen IP-rechten of intellectueel eigendom op rusten. We maken hieronder omwille van het leesgemak gebruik van eerder verschenen materiaal (het QC-protocol, de zogeheten Quicksan QC-Implementatie en opdrachtomschrijving van provincie Gelderland (Von Asmuth, 2018a; Provincie Gelderland, 2022; Von Asmuth, 2022b), her en der met redactionele aanpassingen en updates.

Opslag en uitwisseling van brongegevens

Het belang (dan wel de noodzaak) voor het uitwisselen van brongegevens volgt o.a. uit het QC-protocol van de gezamenlijke provincies en RWS (Von Asmuth, 2018). Belangrijke uitgangspunten daarvan zijn transparantie en reproduceerbaarheid, die op hun beurt vragen om behoud en opslag, en dus ook uitwisseling, van zowel originele gegevens, als uitgevoerde omrekeningen, controles, correcties en keurmerken. Qua gegevensinhoud gaat het bij de grondwaterstandsmetingen per processtap in het bijzonder om:

- Meten: Tijd, waterdruk, luchtdruk (indien relevant)
- Omrekenen: Parameters
- Controleren: Controlemetingen
- Corrigeren: Correcties
- Keuren: QC-label, en QC-status

Voor zowel bronhouders als leveranciers is van belang wat de inhoud, format en methode voor het uitwisselen en opslaan van deze data dient te zijn. Het QC-protocol bevat als richtlijn bovendien het gebruik van, eenduidig en expliciet gedefinieerde dataformats, datadefinities en datamodellen voor opslag en uitwisseling. De genoemde brongegevens vallen echter buiten de gegevenscatalogi van de BRO, en ook andere standaarden zoals die gehanteerd worden door het forum standaardisatie (zie <https://www.forumstandaardisatie.nl>) zijn ontoereikend voor dit doel. Er bestaat op moment van schrijven dus geen formele standaard of gegevenscatalogus voor dergelijke gegevens. Om die reden zijn het bron-datamodel en bron-format voor uniforme data-uitwisseling tussen bron- en QC-applicaties ontwikkeld, dat hieronder nader beschreven wordt. Omdat dit format (nog) niet is aangemeld bij het forum standaardisatie, gaat het echter (nog) niet om een formele standaard en valt daarmee onder het zogeheten 'pas toe of leg uit'-beleid.

Het Bron- en HDF5-format

Het zogeheten bron-format is een variant van het HDF5-format, een open source format dat gebruikt wordt en waarvoor ook open source tooling beschikbaar is in meerdere programmeertalen, inclusief Python en Matlab (zie https://en.wikipedia.org/wiki/Hierarchical_Data_Format). Het hiërarchische karakter van dit format sluit goed aan bij de hiërarchische XML-berichten van de BRO en UML-schema's van de BRO-catalogi. HDF is oorspronkelijk

ontwikkeld door het National Center for Supercomputing Applications, en wordt nu ondersteund door de non-profit organisatie HDF Group (<https://www.hdfgroup.org>). Andere voordelen van dit format zijn:

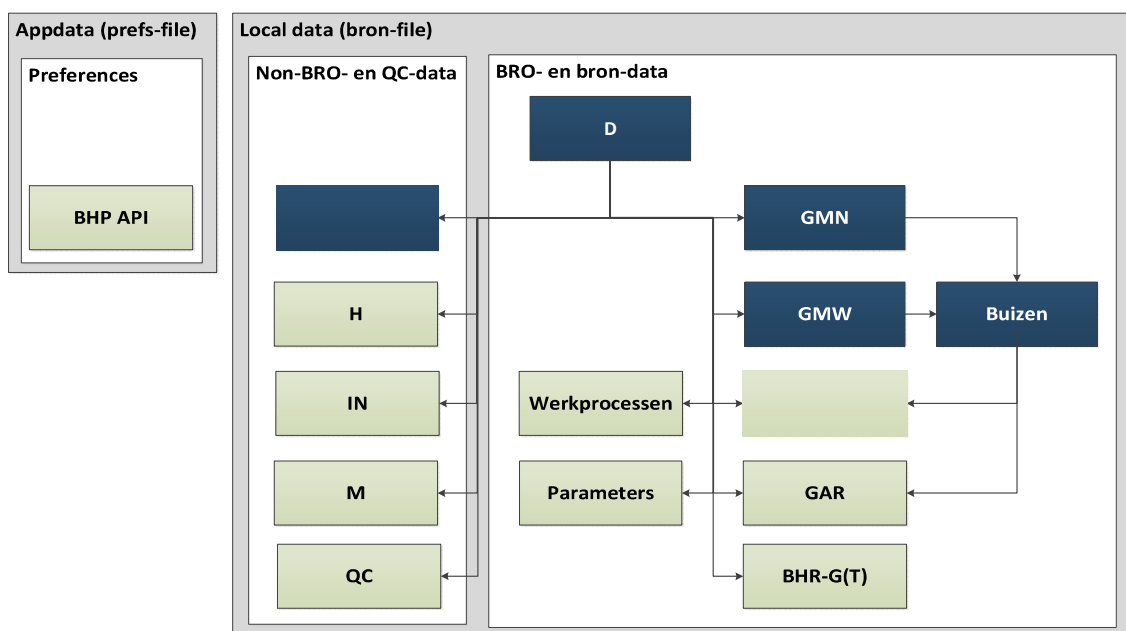
- het dekt zowel de data-inhoud van de BRO, als die van brondata conform QC-protocol én de QC-Wizard volledig af;
- bij gebruik ervan zijn geen verdere import- of export-routines nodig voor data-uitwisseling.

De inhoud van het bron-format is gebaseerd op het spreadsheet- en tabelformat voor BRO-objecten (Von Asmuth, 2020), en de opzet en principes van dit format zijn dusdanig gekozen dat het naadloos aansluit bij de datamodellen van de BRO zelf aan de ene kant, terwijl het aan de andere kant ook ruimte laat voor opslag en uitwisseling van de relevante bron- en QC-gegevens. Dergelijke aanvullende gegevens worden daarbij netjes afgebakend en ‘gemapt’ op de BRO-gegevens. Het spreadsheet- en tabelformat format is bovendien eenvoudig overdraagbaar en vertaalbaar in verschillende varianten en platforms. Het is daarnaast ook het format dat de QC-Wizard zelf gebruikt voor dataopslag en uitwisseling.

Er is zoals gezegd op moment van schrijven nog geen gegevenscatalogus van dit bron-format beschikbaar, maar de opzet en betekenis van de gegevens is zoveel mogelijk ‘self explaining’. De tabellen bevatten, net als het bij HydroMonitor-format (von Asmuth en Vonk, 2017), intern de definitie van de desbetreffende datatypen en/of van toepassing zijnde eenheden en referentieniveaus. Het format bevat ook een verwijzing naar beschikbare, andere documentatie aanwezig (Von Asmuth, 2020; Van Bergen en Von Asmuth, 2022). Er zijn daarnaast ook voorbeeldbestanden en ook een Python-implementatie van beschikbaar op:

<https://github.com/QC-Protocol/BronDatamodel>

De globale opzet en inhoud van het brondatamodel is weergegeven in Figuur 1.



Figuur 1: Schematische weergave van de globale opzet en inhoud van het zogeheten bron-format. Het format valt onder het zogeheten 'pas toe of leg uit'-beleid van forum standaardisatie.

Originele, onbewerkte (druk)sensorbestanden en -gegevens

File	Put	BuisNr	StartDatumTijd	EindDatumTijd	LoggerBrand	LoggerType	LoggerSerial	LoggerDepth	IsCorrupt	IsDuplicaat
[String]	[String]	[Integer]	[ExcelDate]	[ExcelDate]	[Categorical]	[Categorical]	[String]	[m]	[Boolean]	[Boolean]
B26C0046;001_11.06.2010_00.00.00_00.IDC	B26C0046	1	11-6-2010 0:00	13-10-2010 0:00	Keller	Relative	50	2.438	0	0
B26C0046;001_19.02.2010_00.00.00_00.IDC	B26C0046	1	19-2-2010 0:00	10-6-2010 12:00	Keller	Relative	50	2.438	0	0
B26C0046;001_19.09.2009_00.00.00_00.IDC	B26C0046	1	19-9-2009 0:00	18-2-2010 12:00	Keller	Relative	50	2.438	0	0
B26C0046;003_11.06.2010_00.00.00_00.IDC	B26C0046	3	11-6-2010 0:00	13-10-2010 0:00	Keller	Relative	53	2.655	0	0
B26C0046;003_19.02.2010_00.00.00_00.IDC	B26C0046	3	19-2-2010 0:00	10-6-2010 12:00	Keller	Relative	53	2.655	0	0
B26C0046;003_19.09.2009_00.00.00_00.IDC	B26C0046	3	19-9-2009 0:00	18-2-2010 12:00	Keller	Relative	53	2.655	0	0
B26D0005;001_12.06.2010_00.00.00_00.IDC	B26D0005	1	12-6-2010 0:00	13-10-2010 0:00	Keller	Relative	31	2.464	0	0
B26D0005;001_19.09.2009_00.00.00_00.IDC	B26D0005	1	19-9-2009 0:00	24-2-2010 0:00	Keller	Relative	31	2.464	0	0
B26D0005;001_24.02.2010_12.00.00_00.IDC	B26D0005	1	24-2-2010 12:00	11-6-2010 12:00	Keller	Relative	31	2.464	0	0
B26D0005;002_12.06.2010_00.00.00_00.IDC	B26D0005	2	12-6-2010 0:00	13-10-2010 0:00	Keller	Relative	38	2.676	0	0
B26D0005;002_19.09.2009_00.00.00_00.IDC	B26D0005	2	19-9-2009 0:00	24-2-2010 0:00	Keller	Relative	38	2.676	0	0
B26D0005;002_24.02.2010_12.00.00_00.IDC	B26D0005	2	24-2-2010 12:00	11-6-2010 12:00	Keller	Relative	38	2.676	0	0
B26D0090;004_11.06.2010_12.00.00_00.IDC	B26D0090	4	11-6-2010 12:00	13-10-2010 12:00	Keller	Relative	2934	4.947	0	0
B26D0090;004_18.02.2010_12.00.00_00.IDC	B26D0090	4	18-2-2010 12:00	11-6-2010 0:00	Keller	Relative	2934	4.943	0	0
B26D0090;004_19.09.2009_00.00.00_00.IDC	B26D0090	4	19-9-2009 0:00	18-2-2010 0:00	Keller	Relative	2934	4.960	0	0
B31B0107;001_09.02.2010_00.00.00_00.IDC	B31B0107	1	9-2-2010 0:00	22-6-2010 0:00	Keller	Relative	3717	2.995	0	0
B31B0107;001_20.10.2009_00.00.00_00.IDC	B31B0107	1	20-10-2009 0:00	8-2-2010 12:00	Keller	Relative	3717	2.995	0	0
B31B0107;001_23.06.2010_00.00.00_00.IDC	B31B0107	1	23-6-2010 0:00	28-10-2010 0:00	Keller	Relative	3717	2.995	0	0
B31B0116;001_09.02.2010_00.00.00_00.IDC	B31B0116	1	9-2-2010 0:00	22-6-2010 12:00	Keller	Relative	3677	4.938	0	0
B31B0116;001_20.10.2009_00.00.00_00.IDC	B31B0116	1	20-10-2009 0:00	8-2-2010 12:00	Keller	Relative	3677	4.878	0	0
B31B0116;001_23.06.2010_00.00.00_00.IDC	B31B0116	1	23-6-2010 0:00	28-10-2010 12:00	Keller	Relative	3677	4.986	0	0
B31B0126;003_09.02.2010_12.00.00_00.IDC	B31B0126	3	9-2-2010 12:00	22-6-2010 12:00	Keller	Relative	1446	7.025	0	0
B31B0126;003_13.02.2009_12.00.00_00.IDC	B31B0126	3	13-2-2009 12:00	29-7-2009 0:00	Keller	Relative	1446	6.961	0	0
B31B0126;003_23.06.2010_00.00.00_00.IDC	B31B0126	3	23-6-2010 0:00	28-10-2010 0:00	Keller	Relative	1446	6.954	0	0
B31B1003;001_09.02.2010_00.00.00_00.IDC	B31B1003	1	9-2-2010 0:00	22-6-2010 12:00	Keller	Relative	4024	1.494	0	0
B31B1003;001_20.10.2009_00.00.00_00.IDC	B31B1003	1	20-10-2009 0:00	8-2-2010 12:00	Keller	Relative	4024	1.494	0	0
B31B1003;001_23.06.2010_00.00.00_00.IDC	B31B1003	1	23-6-2010 0:00	28-10-2010 0:00	Keller	Relative	4024	1.494	0	0
B31B1003;002_09.02.2010_00.00.00_00.IDC	B31B1003	2	9-2-2010 0:00	22-6-2010 12:00	Keller	Relative	4008	1.488	0	0
B31B1003;002_20.10.2009_00.00.00_00.IDC	B31B1003	2	20-10-2009 0:00	8-2-2010 12:00	Keller	Relative	4008	1.488	0	0
B31B1003;002_23.06.2010_00.00.00_00.IDC	B31B1003	2	23-6-2010 0:00	28-10-2010 0:00	Keller	Relative	4008	1.497	0	0
B31D0126;001_09.02.2010_12.00.00_00.IDC	B31D0126	1	9-2-2010 12:00	24-6-2010 0:00	Keller	Relative	2210	1.503	0	0
B31D0126;001_17.12.2009_00.00.00_00.IDC	B31D0126	1	17-12-2009 0:00	9-2-2010 0:00	Keller	Relative	2210	1.421	0	0
B31D0126;001_24.06.2010_12.00.00_00.IDC	B31D0126	1	24-6-2010 12:00	28-10-2010 12:00	Keller	Relative	2210	1.426	0	0
B31D0126;002_09.02.2010_12.00.00_00.IDC	B31D0126	2	9-2-2010 12:00	24-6-2010 0:00	Keller	Relative	1410	5.111	0	0
B31D0126;002_17.12.2009_00.00.00_00.IDC	B31D0126	2	17-12-2009 0:00	9-2-2010 0:00	Keller	Relative	1410	4.992	0	0
B31D0126;002_24.06.2010_12.00.00_00.IDC	B31D0126	2	24-6-2010 12:00	28-10-2010 12:00	Keller	Relative	1410	5.125	0	0
B31D0134;001_02.07.2010_00.00.00_00.IDC	B31D0134	1	2-7-2010 0:00	1-11-2010 12:00	Keller	Relative	1447	2.488	0	0
B31D0134;001_18.11.2009_00.00.00_00.IDC	B31D0134	1	18-11-2009 0:00	10-2-2010 0:00	Keller	Relative	3679	2.700	0	0
B31D0134;003_02.07.2010_00.00.00_00.IDC	B31D0134	3	2-7-2010 0:00	1-11-2010 12:00	Keller	Relative	1436	2.426	0	0
B31D0134;003_11.02.2010_00.00.00_00.IDC	B31D0134	3	11-2-2010 0:00	1-7-2010 12:00	Keller	Relative	1436	2.355	0	0
B31E0008;001_09.02.2010_00.00.00_00.IDC	B31E0008	1	9-2-2010 0:00	22-6-2010 0:04	Keller	Relative	118	5.110	0	0
B31E0008;001_22.06.2010_12.04.27_00.IDC	B31E0008	1	22-6-2010 12:04	28-10-2010 0:08	Keller	Relative	118	5.110	0	0
B31E0162;001_08.02.2010_12.00.00_00.IDC	B31E0162	1	8-2-2010 12:00	22-6-2010 0:00	Keller	Relative	3721	2.440	0	0

Figuur 2: Voorbeeldoverzicht van sensorgegevens, met bijbehorende relatie met de put (NITGCode of PutCode) en BuisNr waar de metingen gedaan zijn (bron: provincie Utrecht, RHDHV, Trefoil Hydrology (Provincie Gelderland, 2022)).

Er bestaan binnen het grondwaterdomein (naast het bronformat) verschillende formats van verschillende sensor-, data- en/of softwareleveranciers, die ook in huidige praktijk gebruikt worden voor uitwisseling en/of opslag van brondata. Dataloggerbestanden zoals die van Keller en Van Essen (Divers) bevatten doorgaans meer dan de 'Tijdmeetwaardeparen' die in omgerekende, gekeurde en definitieve vorm onderdeel zijn van BRO-registratieobject Grondwaterstandonderzoek (GLD)). Gebruik van dergelijke, originele (druk)sensorbestanden en gegevens kan daarmee een alternatief zijn voor gebruik van het bron-format. De ervaring in dit project (en vele andere) is echter dat dit problemen op kan leveren, doordat data in deze vorm weinig gestructureerd zijn. De bijbehorende relatie met de put waar de metingen gedaan, ontbreekt in dat geval vaak en is uiteraard wel van belang. Het gaat (ook) hierbij niet om formele standaarden, waarmee ze onder het 'pas toe of leg uit'-beleid van forum standaardisatie vallen en qua principe niet voldoen aan het QC-protocol. Dergelijke bestanden voldoen daarnaast ook lang niet altijd aan de QC-protocoleisen op het gebied van data-inhoud en originaliteit (onbewerkte en ongefilterde (druk)metingen). Verdere, bekende aandachtspunten die betrekking hebben op de gegevensinhoud zijn:

- Brongegevens sensoren en loggers (merk, type, bereik, etc.)
- Brongegevens luchtdruk en -compensatie
- Brongegevens omrekening (bekend en/of beschikbaar in de databestanden bij bekende merken zoals Diver en Keller)

Dergelijke bezwaren zijn weliswaar deels oplosbaar, bijv. door de koppeling met de put te verwerken in de naam van de sensorbestanden en door aanpalende gegevens uit te wisselen in een overzichtstabel (*Figuur 2*).