BioDATEN

Ein minimales Metadatenschema für BioDATEN

BioDATEN Abschlussworkshop – Tübingen – 2023

12.06.2023

Überblick

1. Metadaten in der Bioinformatik
2. Minimalschema
   - in Kombination mit anderen Metadatenschema
   - in der Anwendung
3. Diskussion und Ausblick

Bioinformatik

- Methoden aus der Informatik werden auf Fragestellungen der Lebenswissenschaften (Chemie, Biologie, Medizin, Pharmakologie) angewendet
  - Erforschung, Entwicklung und Anwendung computergestützter Methoden zur Beantwortung molekularbiologischer und biomedizinischer Fragestellungen.
  - große Datenmengen auf molekularer und zellbiologischer Ebene
- Gegenstände ihrer Untersuchung:
  - Genome und Gene
  - DNA, RNA, Proteine und wie sie miteinander oder mit anderen Substanzen interagieren
  - physiologische und biochemische Eigenschaften einer Zelle (Replikation, DNA-repair, Proteinsynthese)
  - metabolische Pfade und Netzwerke
  - …
- In welcher Art und Weise beschäftigt sich die Bioinformatik mit ihren Untersuchungsgegenständen. Was versucht die Disziplin an ihnen aufzuklären?
  - Sequenzierung von Genomen
  - Genexpressionsanalyse, um qualitative und quantitative Aussagen über die Aktivität der Gene zu treffen
  - …

Metadaten in der Bioinformatik

- häufig Voraussetzung für die Publikation in einschlägigen Journals
  - Forschungsergebnisse werden in Publikation beschrieben
  - Website des Journals hat Supplementary Information
  - Im Repositorium liegen, z.B. MIAPE-compliant Forschungsdaten
- Einheitlichkeit bei administrativen Metadaten (z.B. DataCite)
- kein übergreifender, einheitlicher, breit verwendeter Standard für fachspezifische Metadaten
- viele Standards für fachspezifische Teildisziplinen

Disziplin-übergreifende Standards

- bioschemas.org
  - baut auf schema.org
  - hat Repräsentationen für
    - ChemicalSubstance
    - Gene
    - MolecularEntity
    - Protein
  - hat aber keine Repräsentationen für Methoden

- ISA framework: Investigation, Study, und Assay (analytische Messungen)
  - kombiniert administrative Metadaten mit fachspezifischen Metadaten
  - Fokus auf Beschreibung von Experimenten
    - Charakteristik von Proben, Technologie und Messmethoden, Beziehungen zwischen Proben und Daten
    - Ziel: Daten und Forschungsergebnisse sollen reproduzierbar und wiederverwendbar sein
  - fast unüberschaubare Anzahl von Werkzeugen (eigenes Ökosystem)

Metadaten-Standards für Teildisziplinen

- Untersuchungsgegenstand, z.B.: Genom
  - Minimal Information about any Sequence Standard (MIxS), mit *checklists, packages, combinations*
  - Eukaryote (MIGSEukaryote), cultured bacteria/archaea, plant (MIGSPlant), virus (MIGSVirus)
- Untersuchungsmethoden (zur ihrer besseren Interpretation und Reproduzierbarkeit)
  - MIAME (Microarray-Experimente)
  - MIAPE (Proteomics-Experimente)
  - MINSEQE (Hochdurchsatz-Sequenzierung-Experimente)

Metadaten in der Bioinformatik

- alle Metadatenformate haben ihr Berechtigung, um Ergebnisse besser interpretieren, einordnen und reproduzieren zu können
- Formate gewährleisten FAIRness innerhalb der Teildisziplin
  - bei entsprechender, gewissenhafter Auszeichnung
  - bei Repositorien-übergreifender Suche (innerhalb der Teildisziplin)
- problematisch bei:
  - Studien, die sich in mehreren Teilfeldern bewegen
  - Datenrepositorien, die Studien aus mehreren Teilfeldern aufnehmen

➢ definiere ein fachspezifisches **Minimalschema** das alle Teilgebiete der Bioinformatik umfasst

➢ zeichne bioinformatische Forschungsdaten fachübergreifend damit aus

  ➢ zusätzlich zu den Teildisziplin-spezifischen Metadaten

Ein Minimalschema für die Bioinformatik

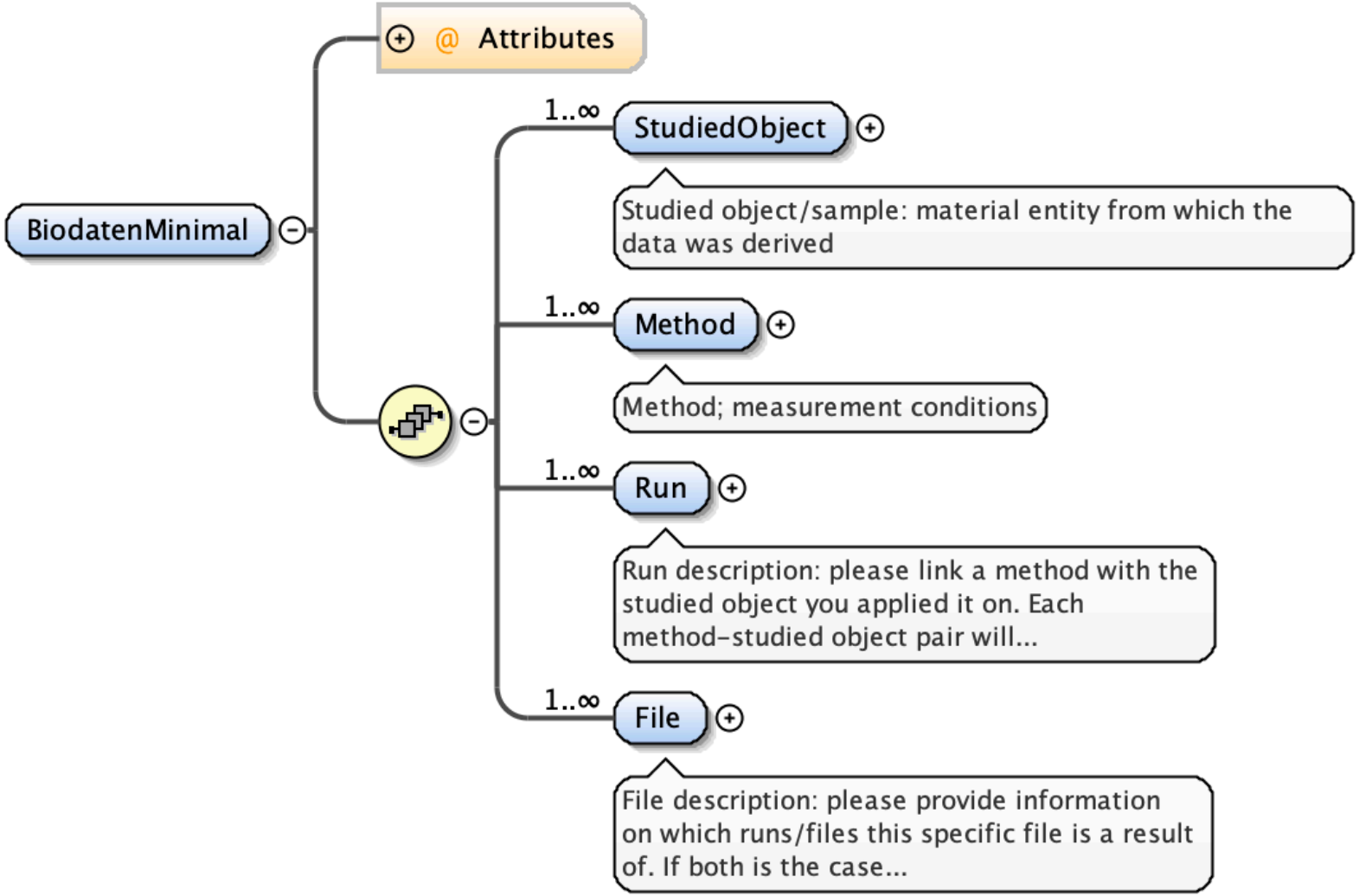
- Anforderungen aller Projektpartner berücksichtigen
  - Forschungsdaten der Partner sollen beschreibbar sein
- minimal, nicht zu wenig, nicht zu viel
  - obligatorische vs. nicht-obligatorische Felder
  - Aufwand bei der Metadatenerfassung gering halten
- zentral, keine *grass-root* Bewegung
  - im Gegenzug, regelmäßige Konsultationen mit Projektpartnern
- basierend auf XML Technologien
  - computer-gestützte Validierung
  - XML-basierte Entwicklung einen Metadaten-Annotations-Tools
- **F**AIR (findability)

CMDI Metadaten-Framework

- Component MetaData Infrastructure
  - ISO Normen ISO 24622-1:2015 und ISO 24622-2:2019
- Metadatenschema werden aus vordefinierten Komponenten/Feldern gebildet
- Fokus auf Wiederverwendung von Definitionen
  - existierende Komponenten: Component Registry
  - existierende Metadatenfelder: Data Category / Concept Registry
- Jede Komponente und jedes Feld hat persistenten Identifikator
- Kann in XSD und XML exportiert werden
- Gewährleistet hohe semantische Interoperabilität falls existierende Komponenten / Felder wiederverwendet werden
- Verwendung von etablierten Ontologien/Thesauri für Feldwerte (via URIs)

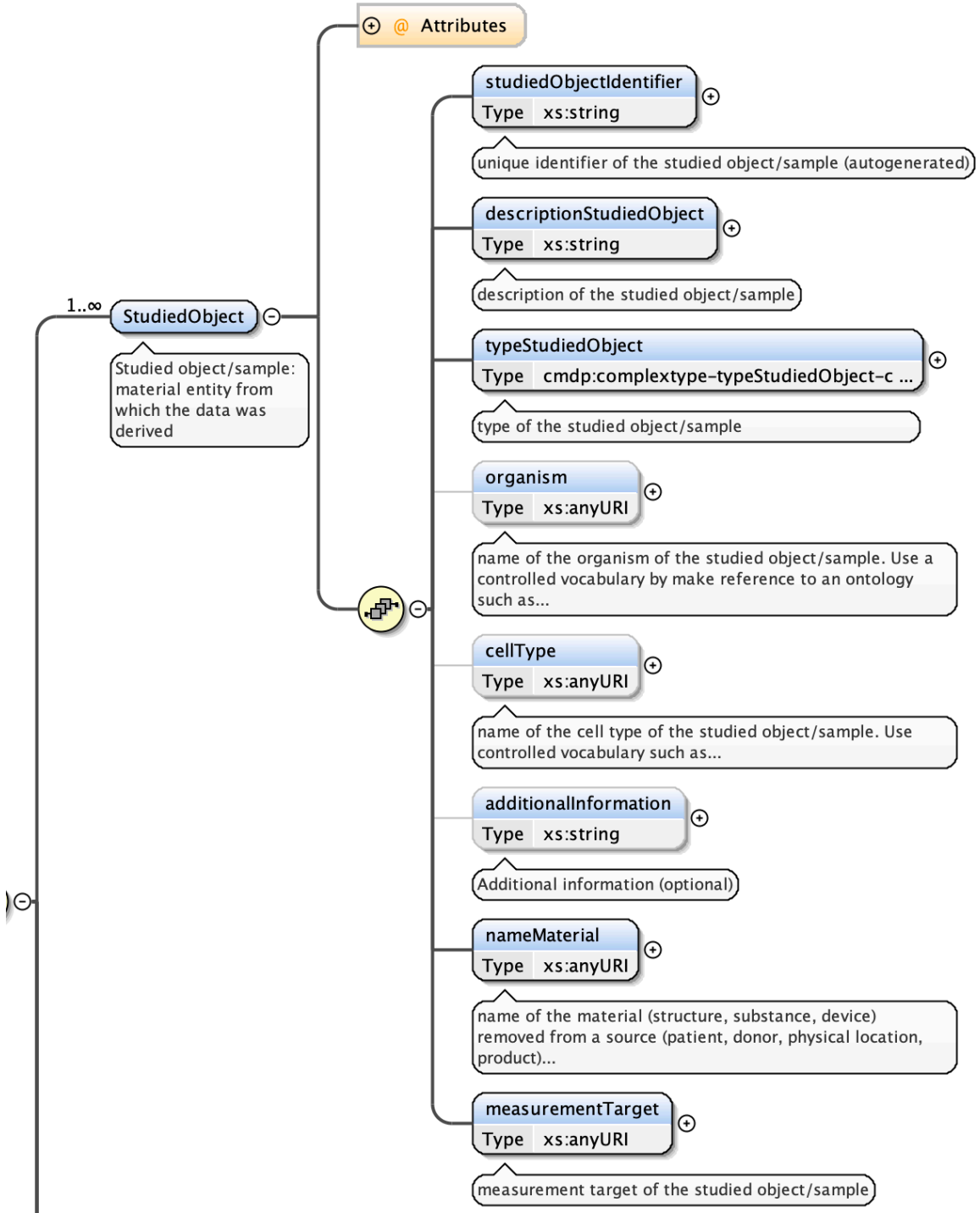
BioDATEN

Minimalschema



BioDATEN

Minimalschema



BioDATEN

Minimalschema

Komponente:

StudiedObject

(Ausschnitt)

Element: **typeStudiedObject**

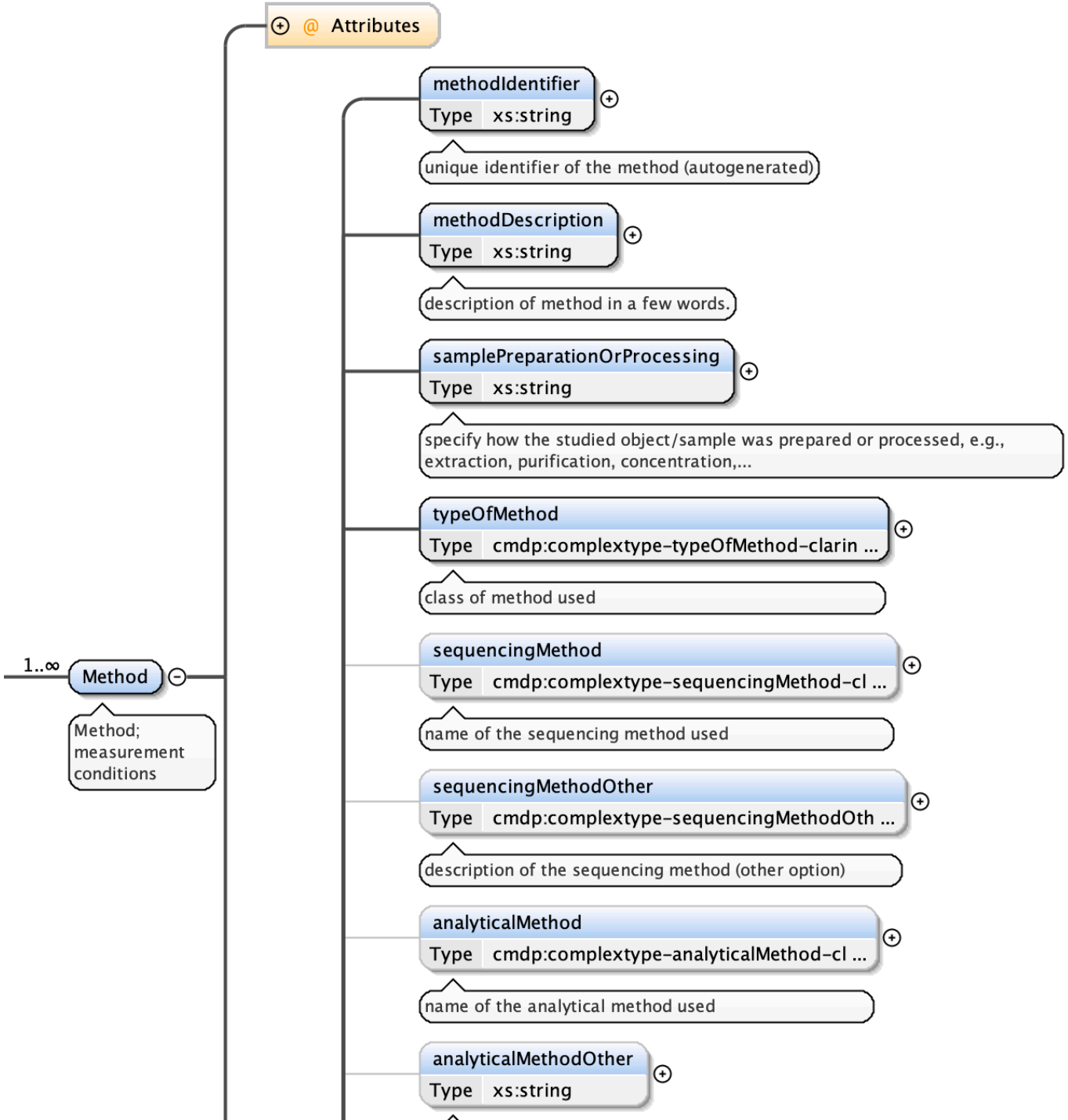
Value scheme:	**Organism (i.e., eukaryota, bacteria, archaea)**
Documentation:	type of the studied object/sample
Number of occurrences:	1 - 1
Cues for tools:	`@cue:dependsOn_valueSelection="'Organism (i.e.,`

Element: **organism**

Value scheme:	**anyURI**
ConceptLink:	http://purl.jp/bio/4/id/200906079942882108
Documentation:	name of the organism of the studied object/sample. Use a controlled vocabulary by make reference to an ontology such as https://bioportal.bioontology.org/ontologies/BERO/
DisplayPriority:	1
Number of occurrences:	0 - 1
Cues for tools:	`@cue:display_placeholder="e.g., Arabidopsis tha`

BioDATEN

Minimalschema



CMDI Metadaten-Framework

Methodentyp:

- *Sequencing method*
- *Analytical method*
- *Protein-protein interaction*
- *Protein-DNA/RNA interaction*
- *Protein-lipid interaction*
- *Particle analysis method*
- *Bioinformatic method*
- *Other*

Analytical method:

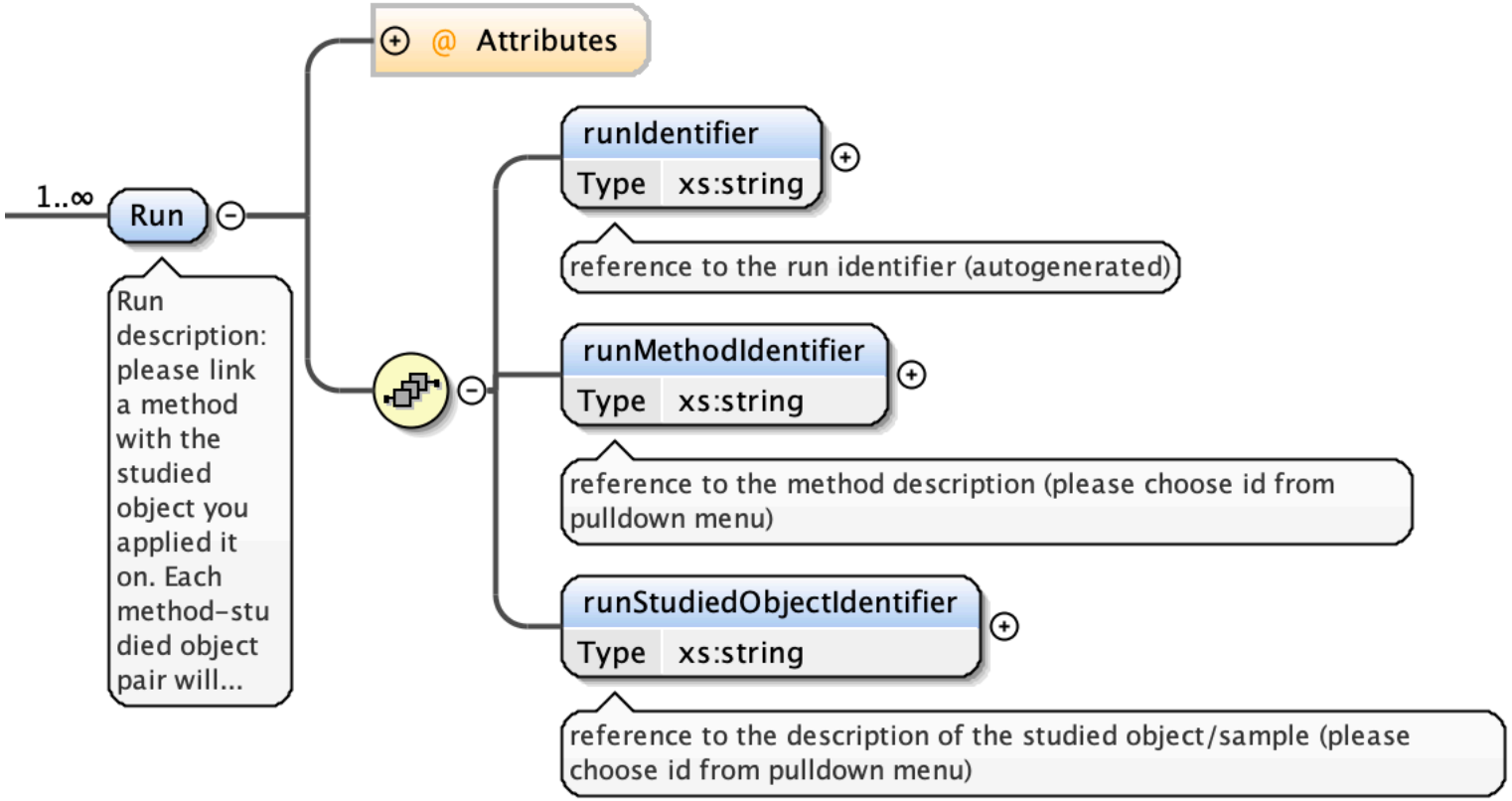
- *High performance liquid chromatography (HPLC)*
- *Gas chromatography (GC)*
- *….*
- *Nuclear magnetic resonance spectroscopy*
- *Matrix-assisted laser desorption/ionization-mass spectrometry (MALDI-MS)*
- *Enzyme-linked immunosorbent assay (ELISA)Combinatorial probe anchor synthesis*
- *Immuno-electron microscopy*
- *Other*

Sequencing method:

- *Maxam–Gilbert sequencing*
- *Chain termination (Sanger sequencing*
- *Ion semiconductor sequencing*
- *…*
- *Helicos single molecule fluorescent sequencing*
- *Microfluidic Systems*
- *Other*

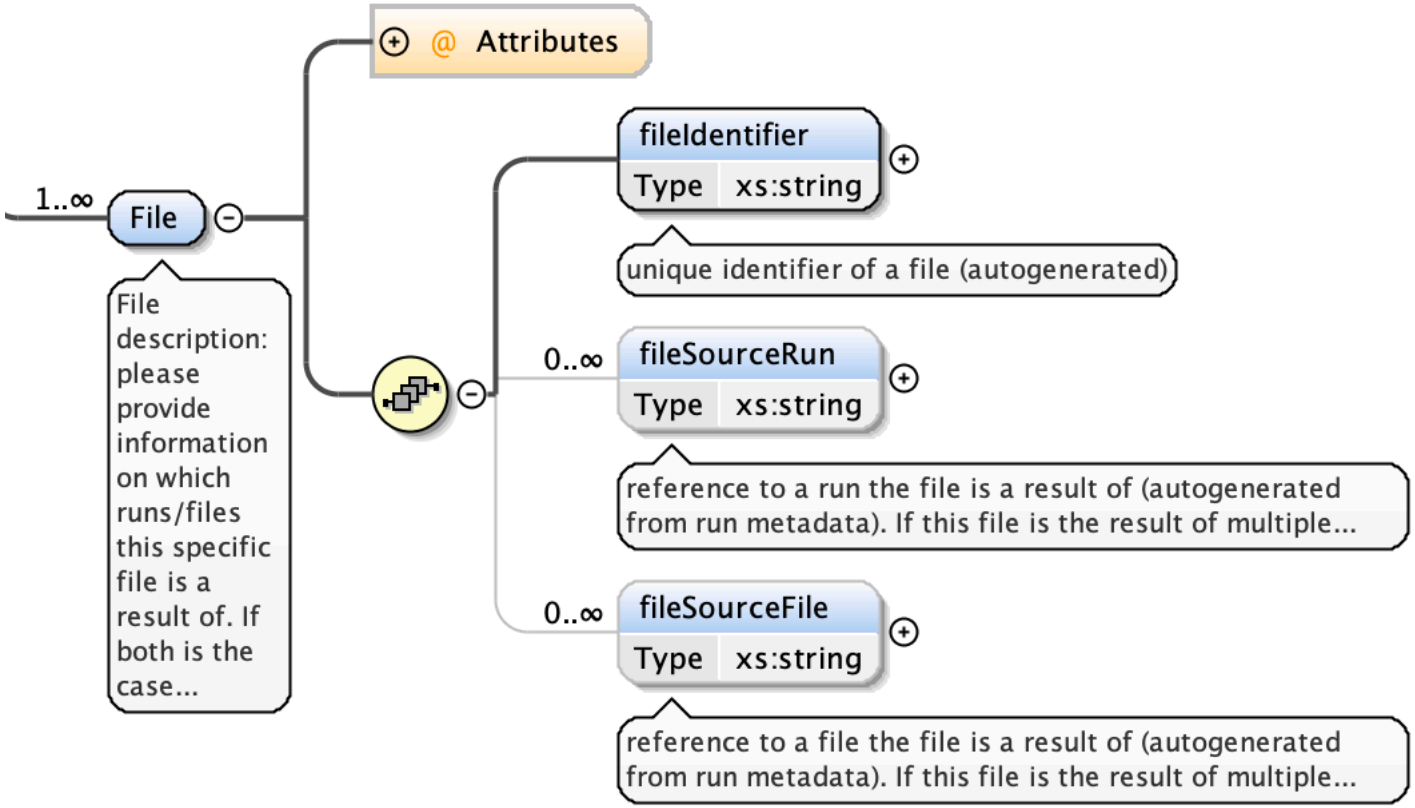
BioDATEN

Minimalschema



BioDATEN

Minimalschema



Proteomic analysis of thioredoxin-targeted proteins in *Escherichia coli*

Jaya K. Kumar, Stanley Tabor, and Charles C. Richardson*

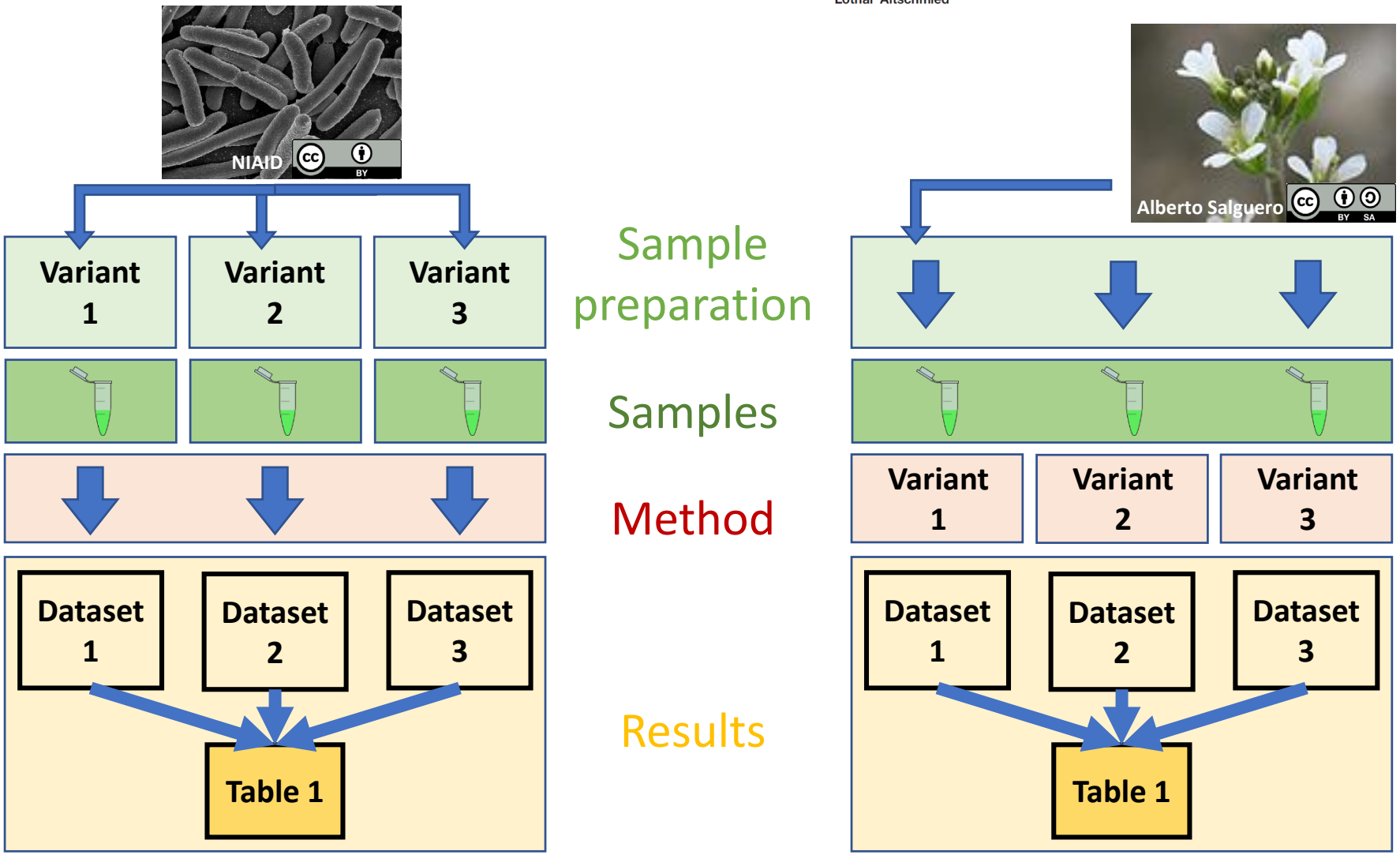
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

Contributed by Charles C. Richardson, December 29, 2003

8240–8254 Nucleic Acids Research, 2012, Vol. 40, No. 17 Published online 22 June 2012
doi:10.1093/nar/gks594

Toward the identification and regulation of the *Arabidopsis thaliana* ABI3 regulon

Gudrun Mönke[1], Michael Seifert[1], Jens Keilwagen[1], Michaela Mohr[2], Ivo Grosse[3], Urs Hähnel[1], Astrid Junker[1], Bernd Weisshaar[4], Udo Conrad[1], Helmut Bäumlein[1,*] and Lothar Altschmied[1]



Discussion & Future Work

- Minimalschema wurde mithilfe von *Use Cases* validiert
- in Kooperation mit den Partnern des BioDATEN Konsortiums entwickelt
- ist verfügbar im öffentlichen Github Repositorium
  - https://github.com/ubtue/BioDATEN-Minimalschema (AGPL-3)
- Idealerweise, das “Dublin Core” für fachspezifische Metadaten innerhalb der Bioinformatik
- Metadaten-Annotationstool unterstützt das Minimalschema
  - optionale Präsentation von Metadatenfeldern
  - Verbindung mit externen Ontologien
  - Top-down-Menus bieten diese kontrollierten Vokabulare als Werte für Felder an
  - ist wesentlicher Bestandteil des Ingest-Prozesses von FD in das BioDATEN Repositorium
  - kapselt fachspezifische & nicht-fachspezifische Metadaten in METS Container
- Aber: fachspezifische Metadaten in Invenio-basierten RDM nur eingeschränkt unterstützt
  - Sicherstellung das über fachspezifisches Vokabular gesucht werden kann (VuFind)
  - Verwendung von DataCite Feld „description“, um möglichst viel fachspezifisches Vokabular unterzubringen