The WheatIS integrated search schema (v1.0)

Table of contents

Changelog

Changes in version 1.0

- No change occurred in v1.0, it is only an version update to approve the usage of the schema across all nodes in the WheatIS federation.

Changes in version 0.2

- Spaces in the field names have been changed to underscores for consistency with the Solr implementation of the schema.
- A new universally unique identifier field ‘id’ has been added to the core type.
- The old, database specific ‘id’ field has been renamed to ‘db_id’.
- The ‘version’ field has been renamed to ‘db_version’ for consistency.
- The type of the ‘version’ field has been changed from int to string.
- The type and definition of the ‘species’ field has been changed from a single integer representing the NCBI taxon identifier to a string (or strings) representing the scientific name(s) of the species associated with the entry.
- The ‘sample’ field has been removed from the core type.
- A generic, multi-valued ‘xref’ field has been added to the core type to allow arbitrary cross-references between sub-types.
- All but the ‘feature_type’ field of the ‘Sequence feature’ sub-type have been changed from required to optional.
- The ‘position’ field has been renamed ‘map_position’ for clarity.
- All the multi-valued fields of the ‘Phenotype’ sub-type have been changed to single valued fields except for the experiment fields.
- The ‘QTL or GWAS result’ sub-type has been split, creating two new sub-types called ‘GWAS result’ and ‘QTL’, increasing the number of named sub-types from eight to nine.
- The definition of the ‘genotype’ field has been changed for clarity.
- An ‘experiment_type’ field has been added to the ‘Experiment’ sub-type.
- Two types of optional ‘xref’ have been added to the ‘Experiment’ sub-type, one for phenotypes and one for accessions.

Introduction

This document defines a conceptual schema for indexing entries across the WheatIS/transPLANT partner databases for the purpose of integrated search. The schema is designed to be simple enough to accomodate results from different and varied sources, yet detailed enough to provide meaningful free text search and subsequent exploration of results via faceting, filtering and rich-snippets.

The schema consists of a core type called a database entry and an extensible number of sub-types. The core type defines a set of fields that we require for any database entry. Each sub-type inherits all the fields of the core type, and can add to or override the definition of core fields, e.g. by adding a sequence position to a sequence feature or by making an optional field mandatory.

The schema is proposed for the purpose of supporting integrated search over different data providers, to give users a single point of entry into multiple databases. As such, the schema isn't intended to be a model of biology or the process of science!

About data types: Currently we define nine sub-types of database entry in addition to the core type, described below. The schema is open for extension if new types of biological entity are proposed for inclusion. All data records are expected to conform to the core 'database entry' type and additionally to be **one** of the other nine sub-types currently defined. Complex records can be modelled across multiple instances.

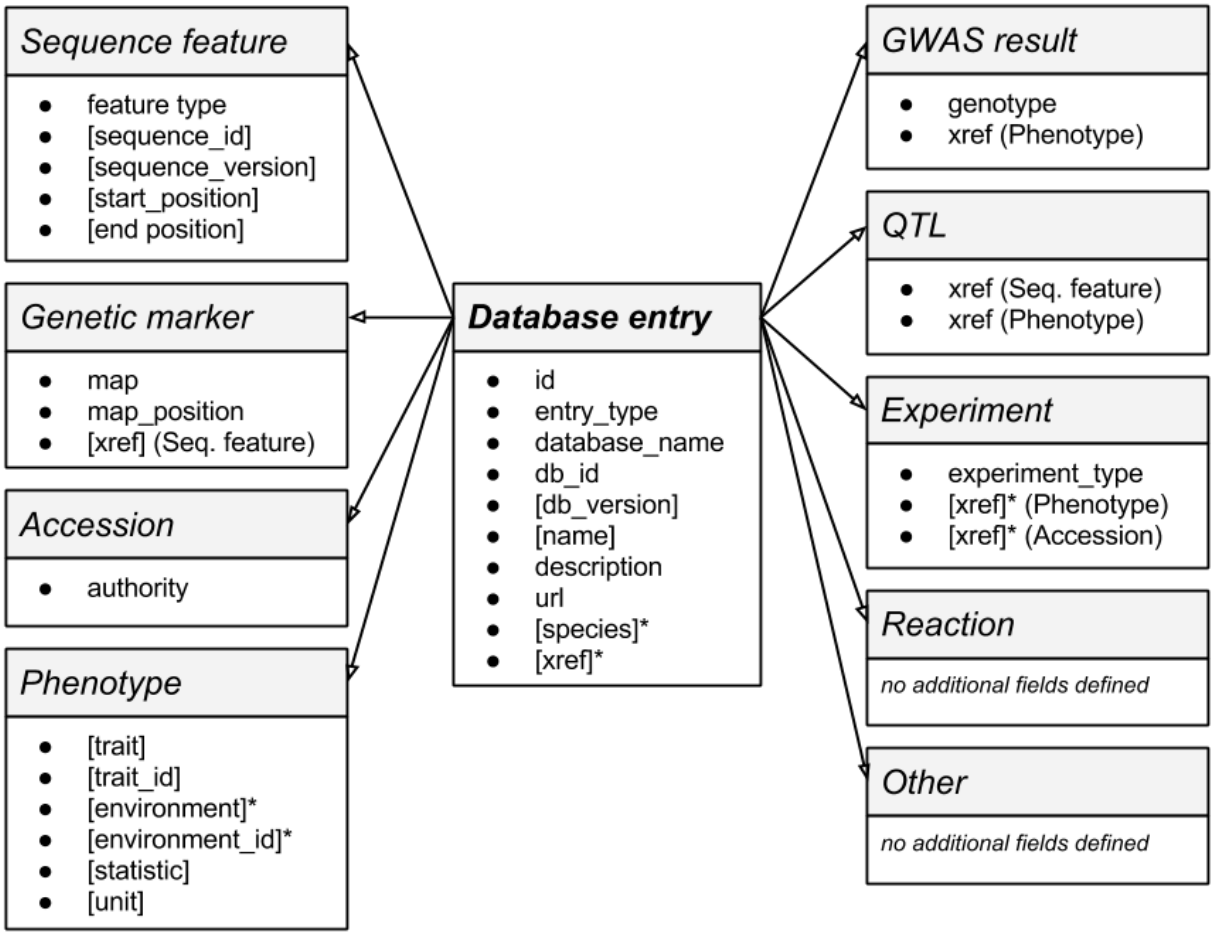
About field types: Any field of type 'text' will be added to the free text search and any field of type 'int', 'float', 'string' or 'enum' is used for faceting, filtering or formatting. Fields are defined as single valued unless otherwise indicated.

This conceptual schema carries its history and the implementation of the underlying Solr schema reflects it by having numerous fields massively unused. It explains the number of fields to represent in the CSV intermediate file.

Definitions of types and fields

Below the core database entry type and the nine sub-types are described and the allowed fields are defined (see Figure 1).

Figure 1, overview of the WheatIS/transPLANT integrated search schema.



Database entry

The core type that all sub-types inherit from. A 'database entry' is defined as any accessioned record describing a biological entity. Fields in bold italic are used for display and search by facets/categories in the WheatIS search tool.

Field	Description	Type	Optional?
id	A universally unique identifier for the entry.	string	Auto generated
entry_type	The type of database entry. Should be one of the sub-types defined below.	string	no
database_name	The name of the source database for the entry.	string	no
db_id	The entry identifier. May be unique for a given source database, at least when concatenated with db_version.	string	no
db_version	Version of the entry in the database, if any.	string	yes
description	Free text description to be indexed for searching. All relevant text related to the entry should be concatenated into this field.	text	no
url	The URL for the entry in the source database.	string	no
species	The scientific name (or names) of the species or other taxonomic classification associated with the entry.	string, multi	yes
xref	A generic cross-reference defined between different documents in the schema. An xref is one or more db_ids prefixed by the sub-type, i.e. "Accession:db_id1" or "Sequence feature:db_id1, Genetic marker:db_id2, ...".	string, multi	yes

Notes on core type fields

ID

The value of this field is auto generated by hashing the concatenation of following other fields:

- database_name
- db_id
- db_version

Entry_type

The entry_type field is not constrained on its value (it is not implemented as an enumeration), but it is highly recommended to use any of following values fitting the best with your data (some entry_type are more generic and can include more specific entry_type). Looking (in the WheatIS portal) at the number of documents matching an entry_type can help you to chose the good one according to your willing to have sparse data filterable by own facet, but which won't be in the top list, or to have you data joining an already existing facet :

- Gene annotation
- Genome annotation
- Physical map feature
- QTL
- Marker
- Accession
- Repeat reference
- Sequence feature
- Bibliography
- Phenotype
- Experiment
- Genetic map
- GWAS analysis
- Transcriptomic gene list

Db_id

Since this field is used to identify the entry in the WheatIS search tool web interface, it is recommended to use a user-friendly name, such as a DOI, a gene name, a phenotyping trait name, what should allow a scientist to rapidly identify the entry among other entries displayed.

Db_version

This field is currently hijacked at URGI to store information allowing unicity of a document, so that we can fill the db_id with a user-friendly name of the entry, the db_id field being displayed on the WheatIS search tool to easily identify a given entry

URL

The URL must be dereferencable so that this backlink to your information system display information on the document.

Species

Species field is not declared mandatory in the Solr schema, but for the purpose of WheatIS federation, it MUST match any of these values in order to be available through the WheatIS search tool, otherwise, the entry will be filtered out:

- Aegilops*
- Hordeum*
- Tritic*
- Wheat*

Other sub types

All other sub-types are unused to this day in the WheatIS search tool, when they will be queried or displayed, this documentation will be updated accordingly.