INDIGO Data Specification

Release 0.1

Open Data Services

Sep 24, 2020

CONTENTS

Table of Contents

ONE

BACKGROUND

This work is part of a new International Network for Data on Impact and Government Outcomes (INDIGO), supported by the University of Oxford's Government Outcomes Lab (GO Lab) See here for a general introduction to INDIGO: <https://golab.bsg.ox.ac.uk/toolkit/INDIGO-Intro2020/>

In early 2020, Open Data Services Co-operative. and the GO Lab collaborated on two initiatives: updating an Impact Bond Database and prototyping the collection of data on returns and non-financial technical assistance in the social investment sector.

These projects were built on the INDIGO Data Specification, an early attempt to standardise the data collected on social impact bonds. The impetus to standardise is familiar from other sectors: a desire to improve data quality, increase transparency and comparability and thereby provide levers for accountability that can improve outcomes and deliver better choices and value for money and in public service delivery.

Openness in the social impact bond sector is not yet the default and, to allay concerns over accuracy and confidentiality, the data model also emphasises traceability and attribution, and the ability to mark fields as disputed or not for public view after collection. Compatibility with these features is built into the accompanying Government Outcomes Lab. application for creating, editing and managing INDIGO data.

THE DATA MODEL

The data model described here is intended as the first steps towards an open data standard. As more data is published using the model, we hope that the ideas here will provoke interest and criticism and be informed by wider conversations in the sector.

2.1 Overview of the data model

A high-level diagram representing the core data model is presented below.
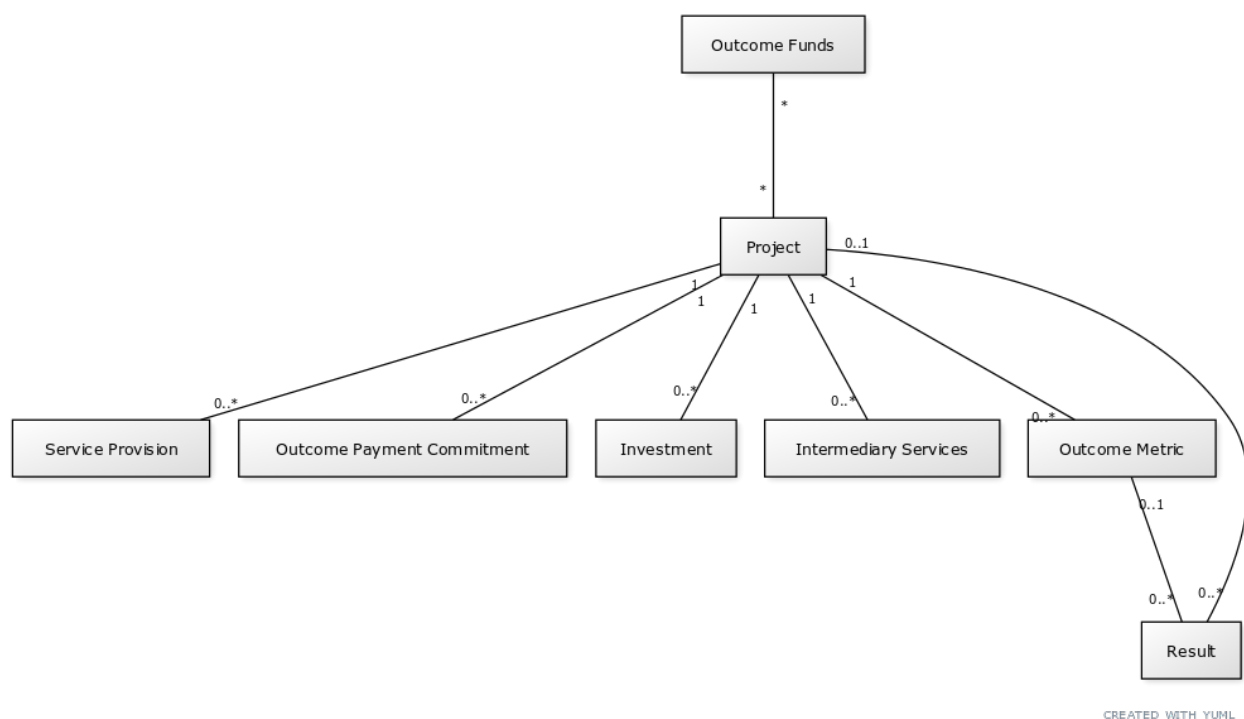
A project is the building block of the INDIGO data specification, representing a single impact bond, social investment or other partnership.

A project can be further broken down into any number of components. These components may represent activities and agreements (like investments, delivery, technical assistance, outcome-based payment agreements and outcome metrics) or measurements of, or associated with, activities and their effect (like outcome metric results or financial performance).

Results and evaluations can be associated either with individual outcome metrics or the project as a whole.

Upstream sources of payment funding can be described by linking any number of outcome funds to a project.

2.2 Project-level and disaggregated data

A cross-sector collaboration is modelled in the INDIGO Data Specification as a project. A project can represent a social impact bond, outcomes based contract or similar partnership.

Project-level attributes provide an overview of a given project, and are designed to provide a useful and unambiguous overview and a backstop in the absence of more granular disclosures.

A project can also be described in terms of its disaggregated components. These individual indicators provide more detail than is available at a project level and are a closer match for the composite nature of how social impact bonds are arranged, delivered and measured.

Project-level data should always be a complete and up-to-date representation of knowledge about a project. Data at a project-level may or may not include and account for the addition of and updates to disaggregated data.

Disaggregated data should be a complete and up-to-date representation of the relevant individual component of a project. Where a corresponding project-level attribute exists, completeness of disaggregated disclosures should not be assumed.

Example

A social impact bond declares total investment committed of USD1,000,000 at the project level. Two disaggregated investment disclosures are available for the project, of USD250,000 and USD500,000. The sum of the disaggregated investments, USD750,000, does not match the total investment commitment for the overall project. There may be several reasons for this discrepancy, including confidentiality, and the project-level figure should be preferred.

2.3 Shared vs project-level data objects

To reduce duplication and error and allow for easier analysis, the INDIGO data standard shares some data across all datasets. The shared objects are: projects, organisations, outcome funds and investment funds.
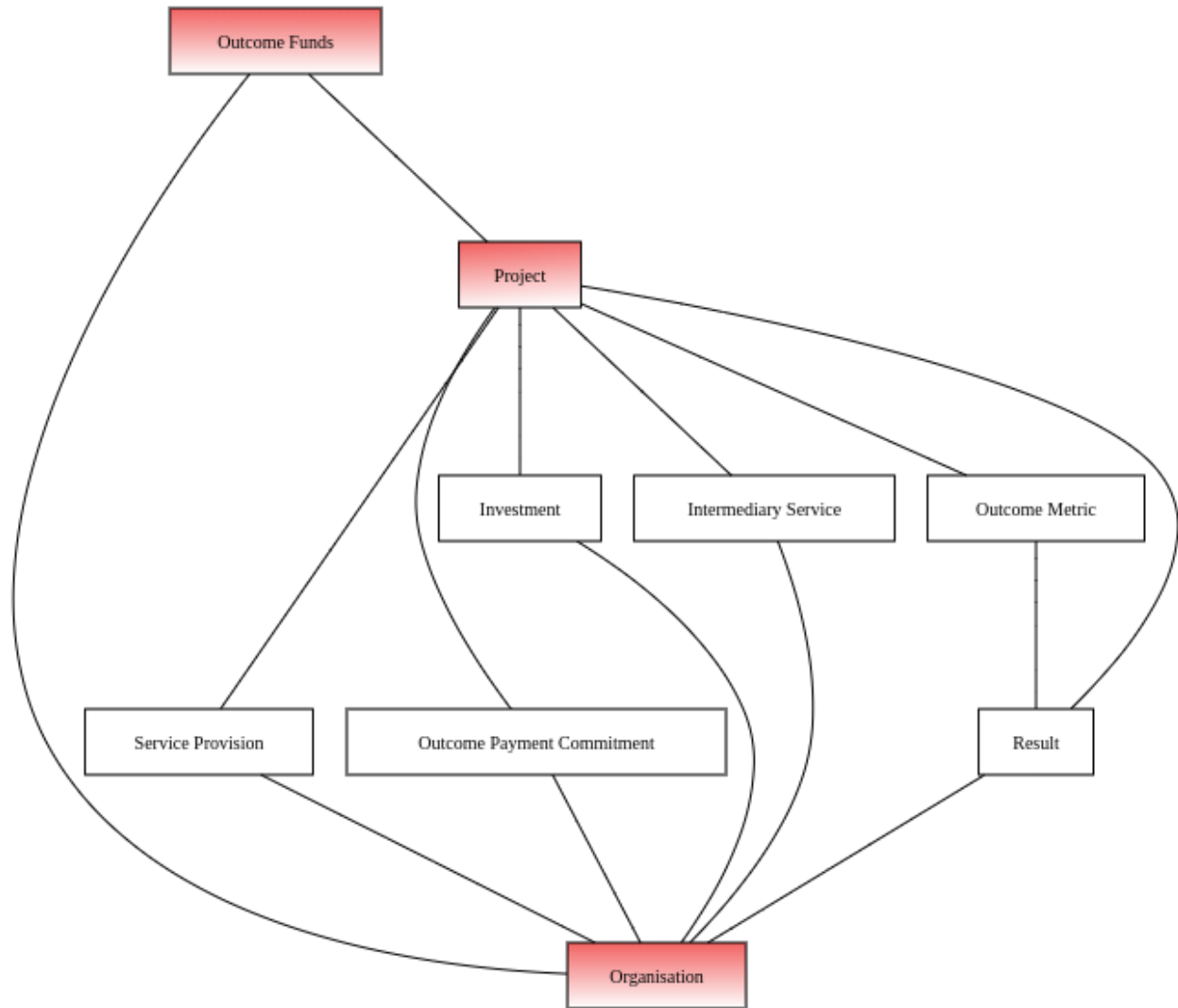
Shared objects are distinguished by identifiers with an 'INDIGO' prefix. See *INDIGO identifiers*.

Each shared object should only be described once, and any new data or corrections added as updates rather than creating a new, duplicate object with a new INDIGO identifier.

2.4 Core and extended datasets

The data dictionary lists the datasets which fields are expected to appear in. The specification has no formal mechanism to restrict the collection of data. But, in general, there is a 'core' set of data fields that all datasets are encouraged to collect; other datasets may collect extra fields for a specific purpose, either for learning purposes or as a result of a particular data-sharing agreement. The core data fields may expand over time, and become more formalised, as the specification matures.

2.5 Organisation roles and classifications

Datasets collected using the INDIGO specification provide two ways to understand organisations.

The first is by using external identifiers, like company or charity numbers, that can be linked to canonical data sources like a company or charity register. These identifiers will allow analysis by organisational type, jurisdiction, sector and other basic demographics. Data sourced *from* these identifiers should not be replicated in the dataset but may be used in analysis.

The second is by allowing a dynamic picture of organisational activity to emerge from the data itself rather than preassigning classifications. This is done by associating individual components (like an investment) of a project with one or more organisations, as in the diagram below.

CREATED WITH YUML
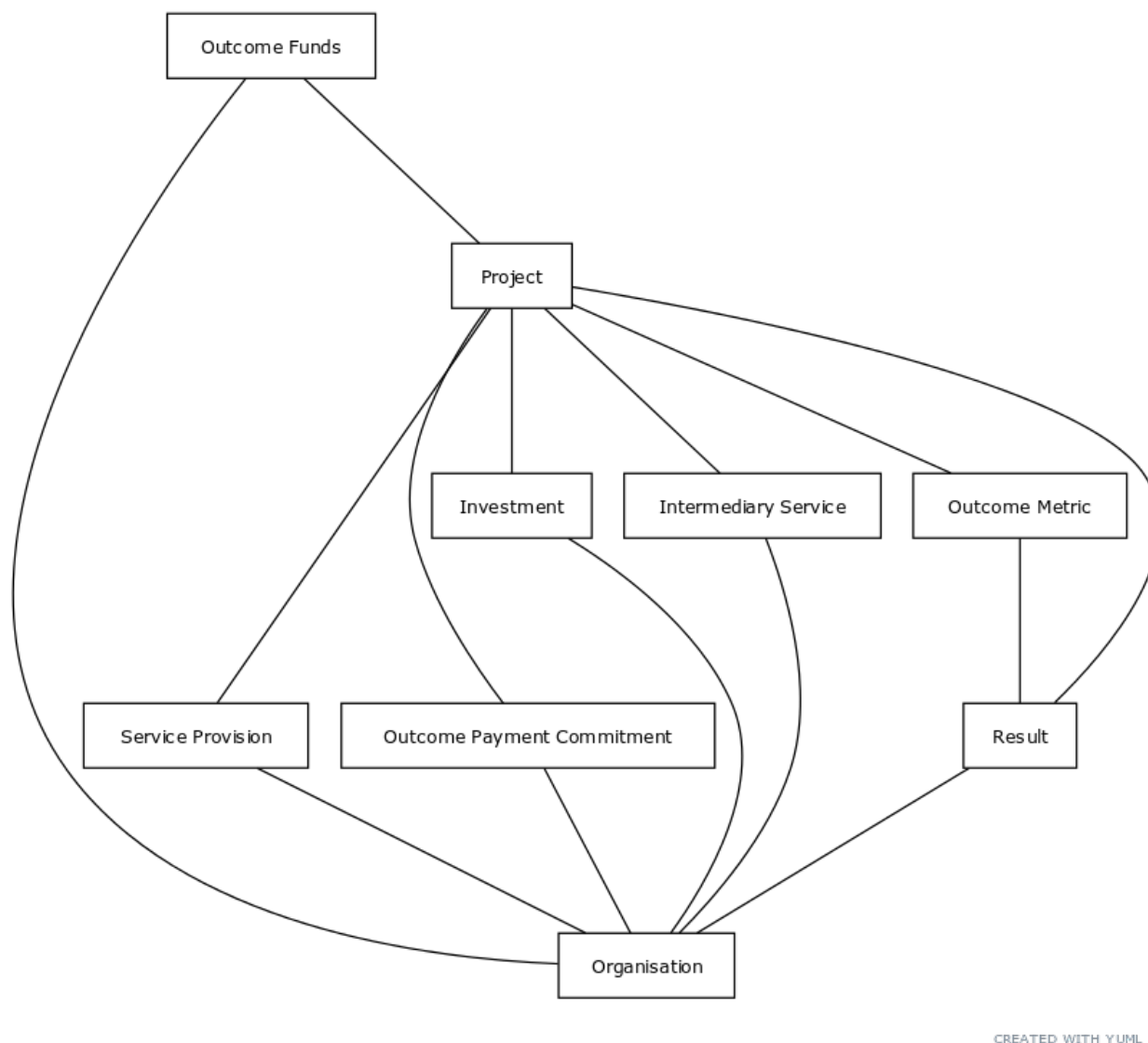
An organisation may appear multiple times in a single project, in different roles, and also recur across projects. The organisation can be traced across projects using *INDIGO identifiers*. Some components also allow an organisation's role to be further broken down via a codelist.

Example

A diversified NGO makes an investment of USD250,000 in a project and is also involved in an aspect of service delivery. The same organisation would appear twice in the data that describes the project: first associated with an *Investment* and then associated with a *Service Provision*. Both of these would describe the organisation in the *Organisational Role Category* of "Nonprofit/NGO".

Data collected in this way allows for analysis of the sector as a whole, as well as changes in organisational behaviour over time.

2.5.1 Relationships between organisations

The data model allows for lightweight modelling of control relationships between organisations, of the form "Organisation A is controlled by Organisation B".

This is useful in cases where, for example, investments are held in subsidiary companies and there is a common parent company that can be used for analysis.

A *controlling* organisation is described like a normal organisation.

A relationship between two organisations is declared by the *controlled* organisation, using the INDIGO ID of the controlling organisation in the *Controlled By* tab.

In general, a relationship between two organisations is assumed to be a relationship between an organisation and its ultimate parent organisation.

WORKING WITH AND COLLECTING DATA

The data specification is described in detail in a data dictionary and in data entry spreadsheets. Three spreadsheet templates are provided, for projects, organisations and outcome funds.

Advanced users may wish to consult the JSON Schema used to transfer data from spreadsheets to the database application.

> **Warning:** The JSON Schema describes the structure of the data model but not data types.

3.1 Data dictionary

On August 27, 2020 GO Lab released a draft data dictionary for participants of our INDIGO Hack-and-Learn event: <https://s3.eu-west-2.amazonaws.com/golab.prod/documents/DataDictionary2020_v2-00_Draft_27-AUG-20.pdf>

For each variable, the data dictionary lists: - a name; - a definition; - the data type; - any technical notes on the data; - the datasets in which the variable can appear; - the status, source and history of the variable.

The data dictionary is currently available as a **Word file**.

3.2 Data collection spreadsheets

Cell colour	Rule
Red	Cell not editable.
Orange	Editable cell; value taken from elsewhere.
Green	Id field; must not be changed after creation.
Grey	Editable field not used elsewhere.

3.3 Identifiers

The INDIGO specification uses three kinds of identifier to link data internally and offers space to enhance the data with the identifiers of related disclosures.

3.3.1 INDIGO identifiers

INDIGO identifiers are assigned to projects, organisations and funds to ensure uniqueness for these important entities across all published datasets. An INDIGO identifier must not be changed once assigned.

The entity an INDIGO identifier refers to can be inferred from the prefix, as follows.

Prefix	Entity type
INDIGO-POJ	A project.
INDIGO-ORG	An organisation.
INDIGO-FUND	An outcome payment or investment fund.

3.3.2 Real-world identifiers

Most organisations will have an official registration number suitable for use as a unique identifier. The INDIGO specification requires identifiers to use the format and prefixes specified by org-id, an open register of organization lists.

An organisation identifier consists of:

1. A list code: a prefix that describes the list the identifier is taken from.

2. An identifier taken from that list.

Example

Open Data Services Co-operative Limited is a private company limited by shares, registered in the UK. From the org-id page the prefix for Companies House is GB-COH. From the linked register the company number is 09506232. The full identifier in org-id format is then GB-COH-09506232.

3.3.3 Internal identifiers

Internal identifiers are unique within a project and used to join components of a project together, for example a result can be linked to a specific outcome metric. Once an internal identifier is set it must not be changed.

3.3.4 Related contract identifiers

To link one or more contracting processes published to the Open Contracting Data Standard (OCDS), use the *ocid*, or contract processing identifier, field. The value in the INDIGO dataset must match that in the relevant published OCDS field. The use of this field is described in the OCDS documentation. The data dictionary describes in what circumstances a contracting process is considered to be linked to a project.

3.3.5 Related grant identifiers

To link one or more grants published to the 360Giving Data Standard, use the *grant_id*, or grant ID, field. The value of the *grant_id* field in the INDIGO dataset must match that in the relevant 360Giving field. The use of this field in is described in the 360Giving documentation. The data dictionary describes in what circumstances a grant is considered to be linked to a project.

SOCIAL INVESTMENT PROTOTYPE

The Social Investment Prototype offers additional tabs to describe technical assistance, grants and the financial aspects of projects in greater detail, including individual transactions.

4.1 Investment Details

Expected and latest internal rates of return can be recorded on the investment details tab.

4.2 Grants

The grants tab allows for a fuller description of grants awarded as part of a project.

The fields in this tab are reused from 360Giving. It is highly recommended that:

- If a grant is already published to the 360Giving Standard then only the *360 Giving Identifier* field should be filled out. This avoids data entry error elsewhere and allows for the possibility of updates to the source data.

- Users should refer to the 360Giving documentation. for detailed guidance on field usage.

An additional field *purpose* is used to indicate what the grant is used for. For compatibility, this shares entries with a similar codelist used in the Social Economy Data Lab specification but adds further entries used in the wider sector.

4.3 Transactions

The transactions tab is designed as a ledger of money in and money out of a project.

A transaction is modelled with a sending organisation and a receiving organisation, a date and an amount. These fields are required.

The value of a transaction (*Amount*) must be positive.

A transaction can be linked to the project as a whole (the default) or to a:

- Outcome payment (using the Outcome Metric ID column to link to the relevant row on the Outcome Metrics tab);

- Investment (using the Investment ID column to link to the relevant row on the Investment tab); or,

- Grant (using the Grant ID column to link to the relevant row on the Grants tab).

Only **one** of these IDs should appear per row, i.e. transactions should be disaggregated where possible. This is particularly important if the data is to be used in further analysis or visualisations.

The transaction type field is used to identify the purpose of the transaction.

The formatting rules on dates and currency values should be followed.

4.4 Technical Assistance

Technical Assistance is modelled as a period of engagement between a funding organisation and a recipient organisation with a defined start and end date.

This period of engagement can be broken down on the Technical Assistance Details tab into a series of component activities (as well as high-level information that covers the whole engagement where appropriate).

An activity is linked to an engagement by using the relevant *id* from the *Technical Assistance* tab. Multiple activities can be linked to a single engagement.

Data should not be provided if no technical assistance from a given category has been given, i.e., zero values are not necessary but assumed in the absence of data.

Each activity can be assigned a cost and a cost type and a time cost in days.

Where no cost is available, or where a cost is inappropriate, a row can be added with the relevant category selected and the value and cost type fields left blank. This will allow activities to be analysed as simple counts.

Activities can be further classified by the delivery approach.

4.5 Formatting data

4.5.1 Dates

The specification allows for imprecise dates depending on how much information is known (e.g., 2020 or 2020-06). Dates must use the YYYY-MM-DD format.

4.5.2 Currencies and currency conversion

A field describing a monetary value in the INDIGO specification should have an accompanying currency field. Monetary values must be described as numbers only with no currency symbols, commas or textual descriptions of large numbers.

Currency codes must come from the ISO 4217 code list.

Example

A social impact bond makes an investment of USD250000. This data should be input as:

Field	Entity value
investment_commitment/currency/value	USD
investment_commitment/amount/exact/value	250000

Inputting the value as "250,000", "$250000" or "250k" would be wrong.

Monetary values should be input in the currency of the original transaction. There may be a converted USD value of any transaction, calculated by the INDIGO project, using the methodology described in the data dictionary. Data providers should not convert transactions to USD when supplying data.

4.6 About

This data model was produced by Open Data Services Co-operative. as part of a project with the University of Oxford's Government Outcomes Lab.