Dimensions

Dimensions as data source: imagine, create, implement

ICTeSSH 2020

Cristina Huidiu
Product Specialist, Dimensions
c.huidiu@digital-science.com

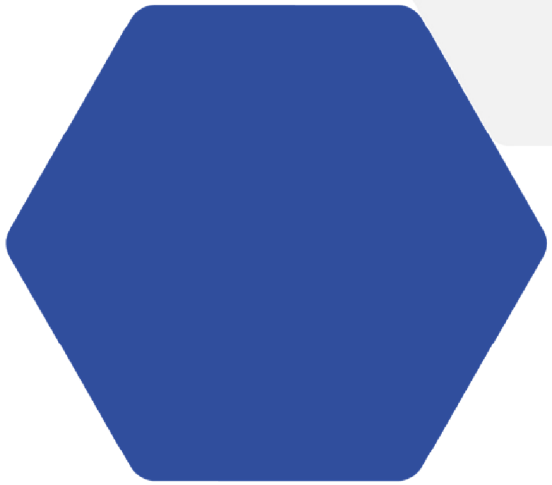
Part of DIGITAL science

Agenda

What is Dimensions?
What data is in Dimensions?
Disambiguate external organisations

Break - 10 minutes

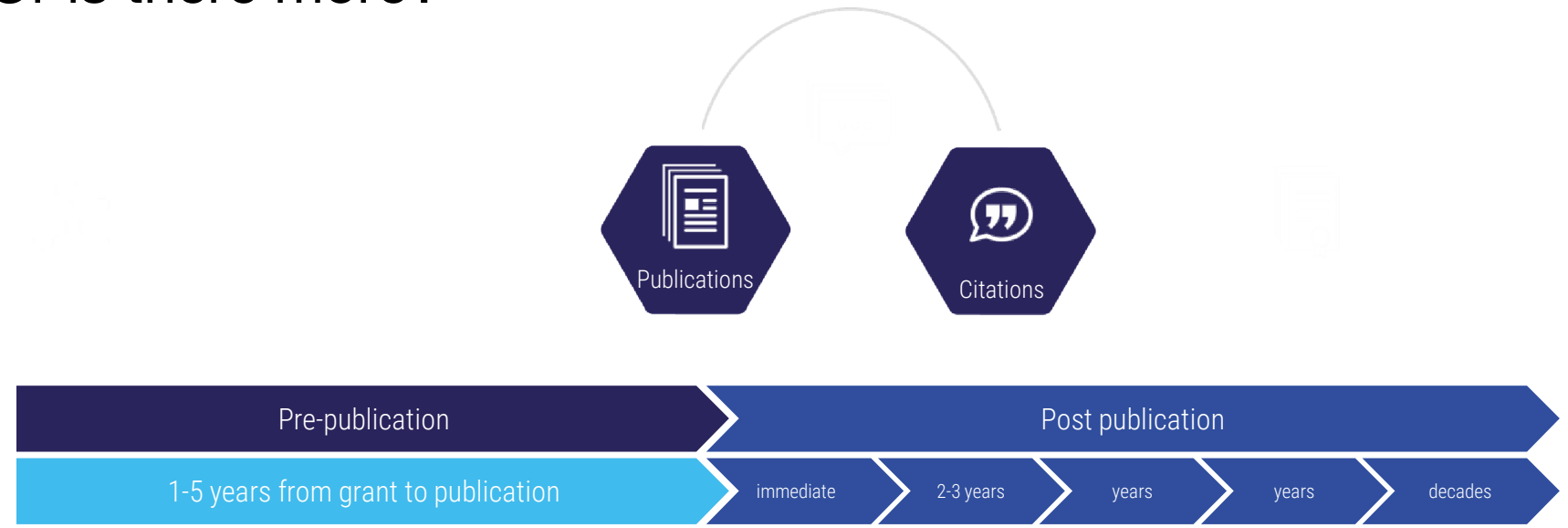
Special functions:
- -classify
- -extract concepts
- -extract grants



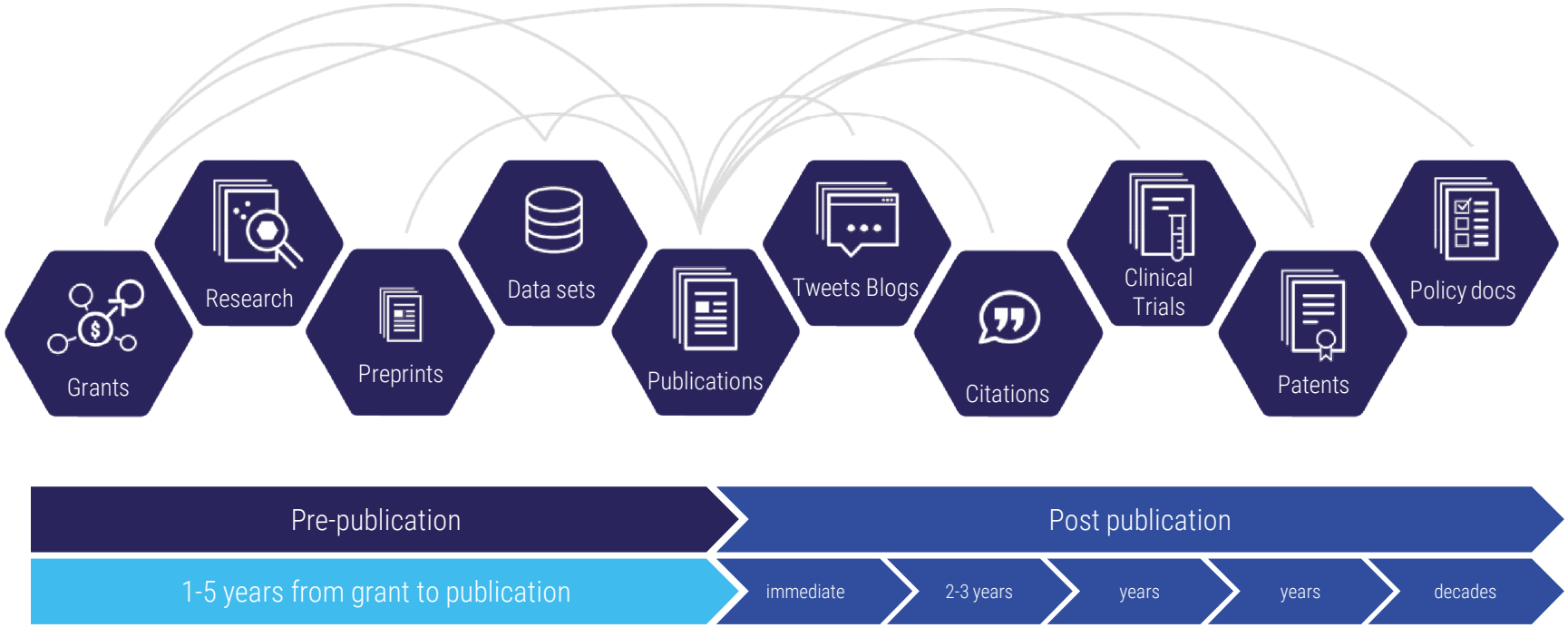
Are publications and citations sufficient for discovery?

Or for research impact assessment?

Or is there more?



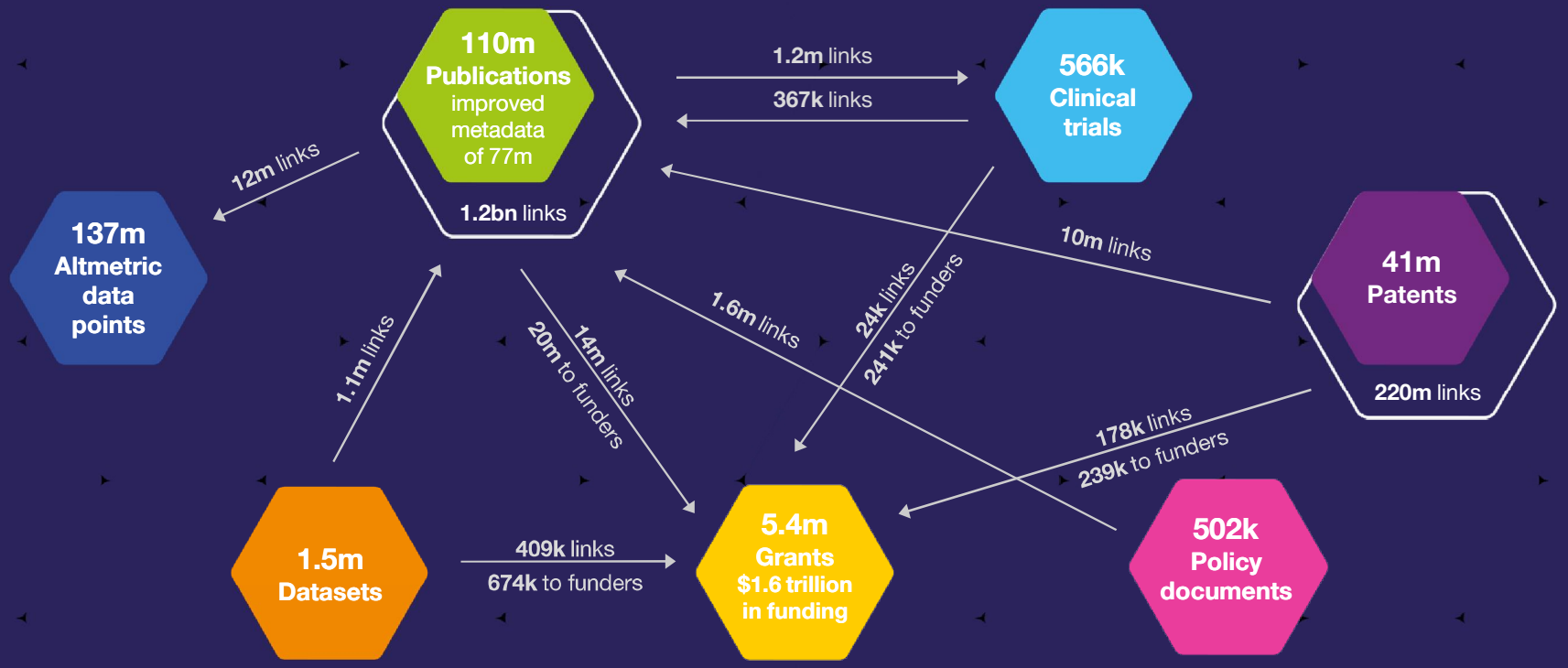
A much broader information landscape is available!



Dimensions

Part of DIGITAL science

… and how that looks in Dimensions



Dimensions

Status: June 2020

Part of **DIGITAL** science

Faster discovery, with more context

A major goal of Dimensions was to create the most **comprehensive** publications database, updated as close as possible to **realtime**, while delivering publications in **context**.

- 110 million publications and counting
- Updated daily
- Full text indexing for 70% of publications
- Citations, references and online attention, all directly accessible.
- Article-level document classification using AI approaches
- Links to grants, funders, citing patents, policy documents, and related clinical trials.

Comprehensive inclusion, delivered quickly and in context.

Dimensions

Part of **DIGITAL** science

Discovery Funded grants data provides *earlier* discovery

With other major data types, *discovery no longer has to be limited to publications*.

Who is working on Coronavirus research projects right now?

- **Over 5 million research grants** from more than 500 funders worldwide
- Explorable connections to resulting publications
- Globally there are **102 research grants working on coronaviruses *right now***. Dimensions can tell you where they are, who they are, what they are doing and who funded the work.

Funded grants make research discoverable before it appears in publications

Dimensions

Part of **DIGITAL** science

Patents, clinical trials, datasets, policy documents and altmetrics give the full picture

- Over 40m patents
- Over 500,000 clinical trials
- 1.5m datasets
- 500,000 policy documents
- Altmetric data for immediate attention

All documents are linked where relevant and all data sources are searched simultaneously. No matter the question, Dimensions can provide a view to find the answer……...

Innovation analysis? Publication links to patents

Immediate impact? Altmetric attention globally

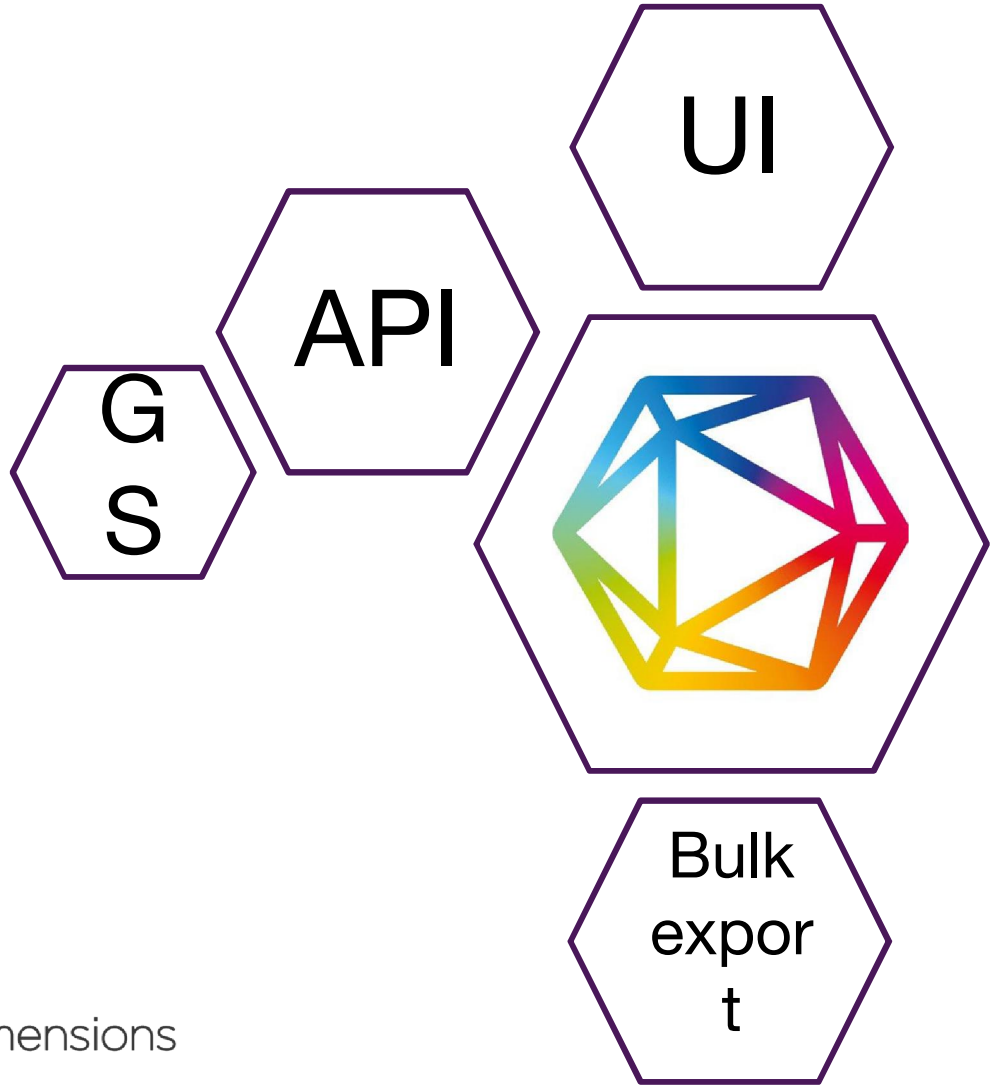
Long term societal impact? Policy documents and clinical trials

Understand the wider picture with patents, clinical trials, datasets policy documents and altmetric attention

How is all the data accessible?

Dimensions

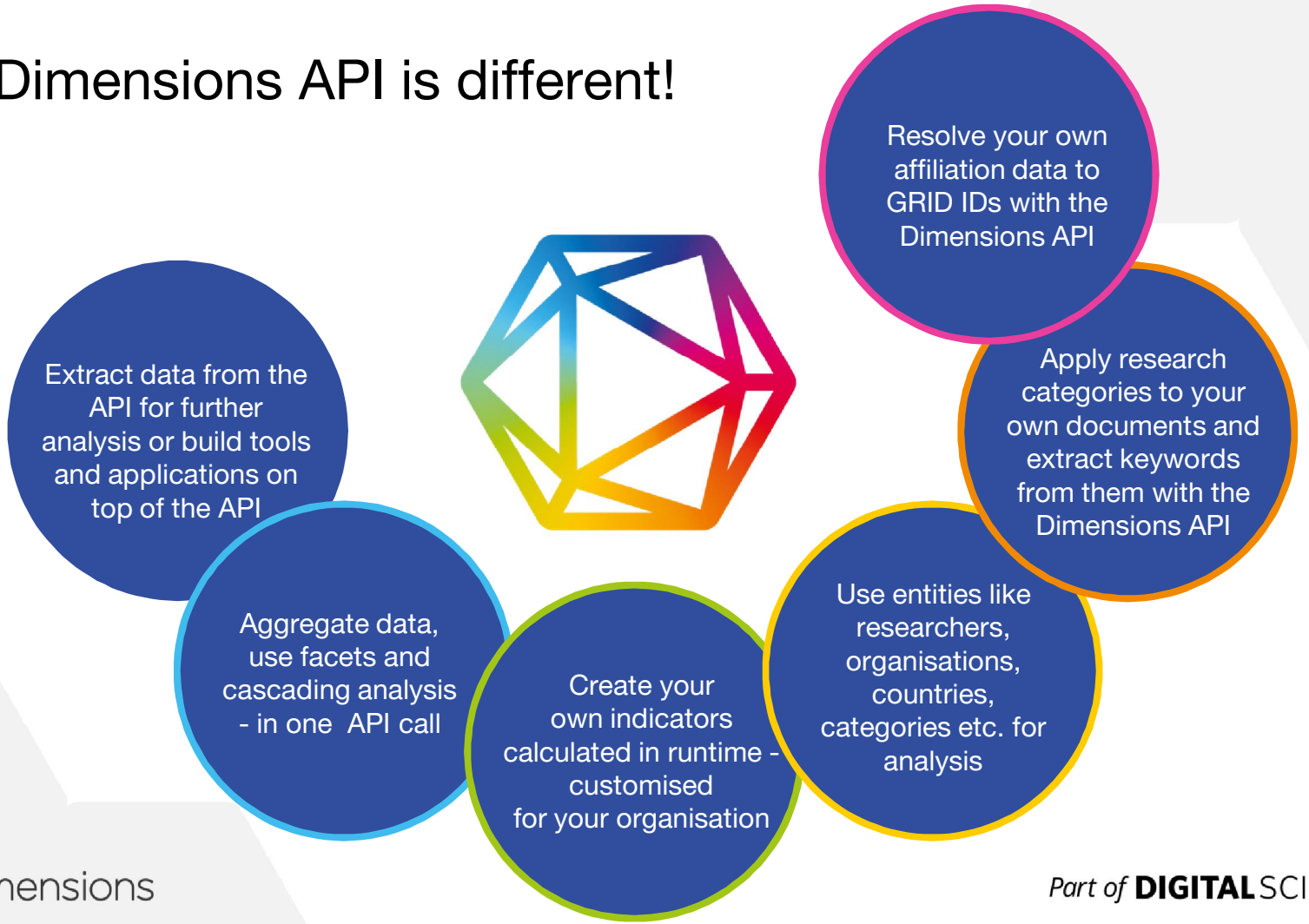
Part of DIGITAL science



Dimensions

Part of **DIGITAL** science

The Dimensions API is different!



Dimensions

Part of **DIGITAL** science

Dimensions

Disambiguate external organisations

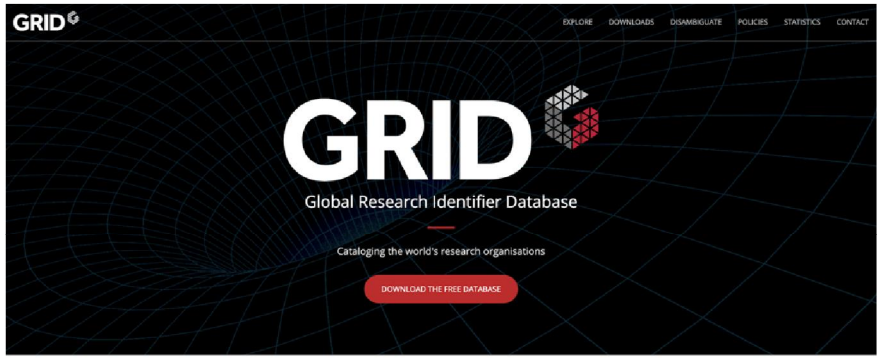
How Dimensions can help to clean and maintain external organisation data

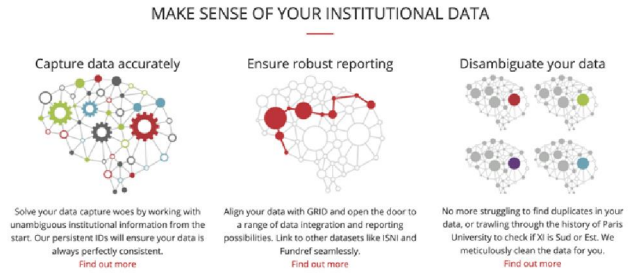
Part of DIGITAL science

The challenges around external organisations

- (Unmanaged) organisation records usually flow into the CRIS/RIM systems when importing data or when being keyed in as supporting/supplement data
- Typically the data quality is not high and error-prone resulting in different names, organisations from different hierarchy levels, typos, encoding issues, etc. or simply duplicates
- As a result cluttered, unstructured information accumulates over time and becomes a burden; e.g. prevents accurate reporting on collaboration reporting
- Generally, the information is captured, but needs cleansing to make sense of it

Starting 2015 when Digital Science launched GRID as an open resource





CC 0 PD

Coverage

Institutes ...	Number	Relative
... with address	97819	100 %
... with type	95605	97 %
... with geographic coordinates	88685	90 %
... with URL	88196	90 %
... with Wikipedia URL	88196	90 %

Dimensions

Part of **DIGITAL** science

In the past: Affiliation resolution as an internal process

Dimensions Enrichment Backend

Unstructured affiliations:

Institute of Physics II, University of Cologne, Cologne, Germany

Structured information:

- **Name**: University of Cologne
- **City**: Cologne
- **Country**: Germany

Name variants like "Universität zu Köln, Cologne University, University to Cologne, University of Cologne, Köln University"

Dimensions Matcher

grid.6190.e

Dimensions

Part of DIGITAL science

Today: Affiliation resolution for everybody

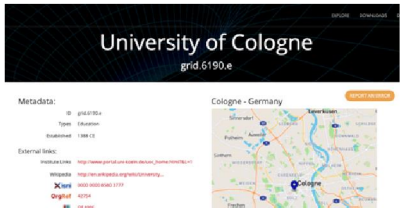
Unstructured affiliations:

Institute of Physics II,
University of Cologne,
Cologne, Germany

Structured information:

- **Name**: University of Cologne
- **City**: Cologne
- **Country**: Germany

Dimensions API



Dimensions

Part of **DIGITAL** science

We ran a pilot with to AAU and DTU of the OPERA project

On one occasion, there was the same research organisation identified over 350 times!

- Between **53%** and **61%** of external organisations could be matched to GRID records
- Between **32%** and **52% duplicates** could be identified
- Potentially the amount of external organisation could be reduced by **13,958** and in another case by **14,861** records and do no longer need maintenance or pollute the system.
- Improved matching with more context:
  - Adding **country**: Matching improved by **20%**
  - Adding **city**: Matching improved by **31%**

How well does it work?

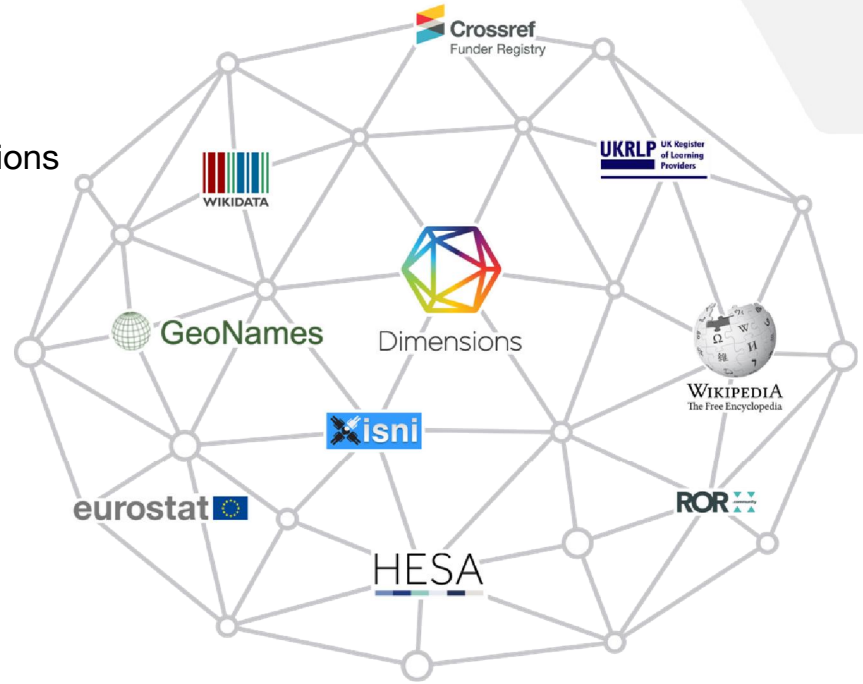
Dimensions

Part of DIGITAL science

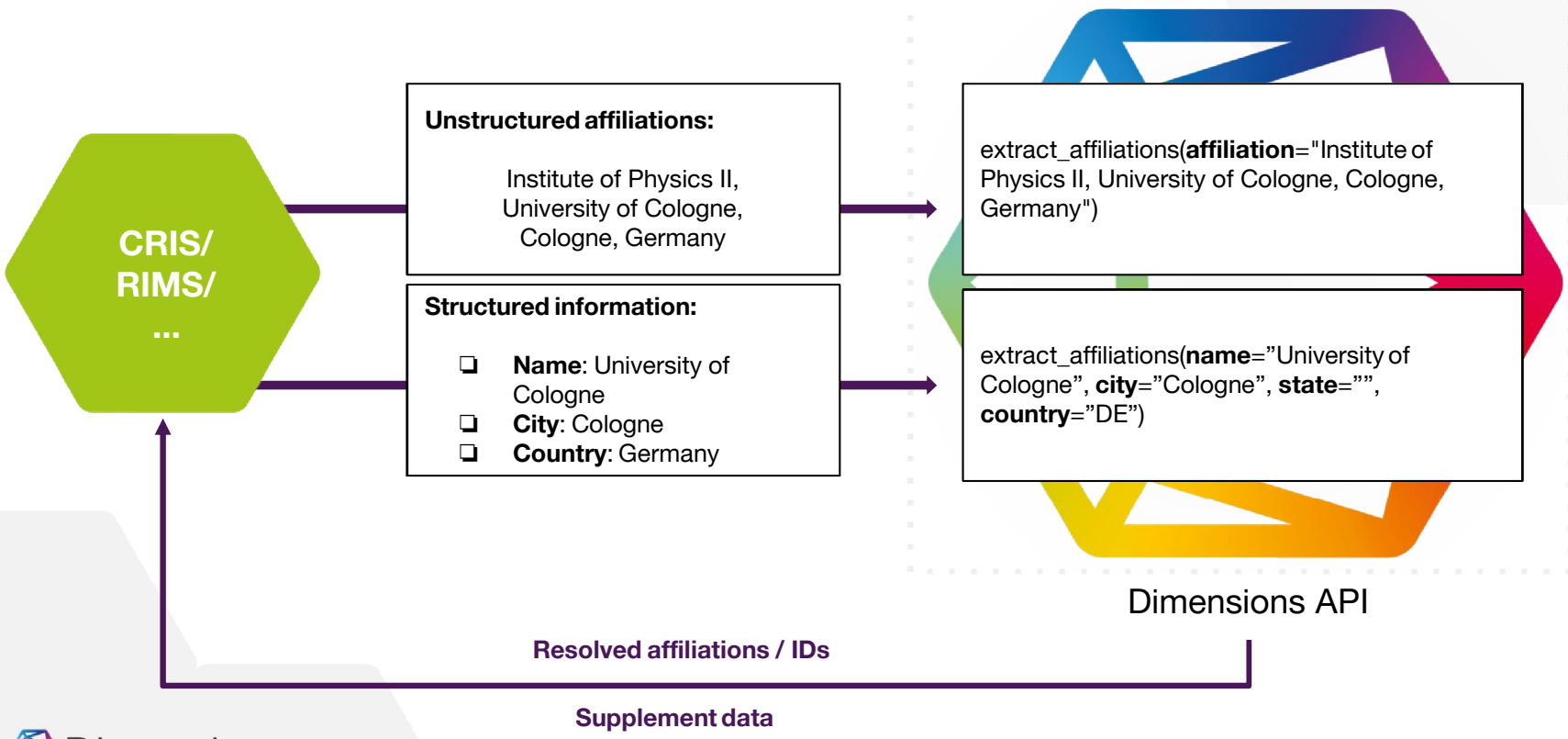
Disambiguate: Additional perks

With the GRID data being available over the Dimensions API, the data can be used to import supplement metadata:

- **Location**: Latitude and longitude
- **Type** of the organisation
- **Hierarchies** and **relations** to other organisations (parents, childs, related organisations)
- **City** and **country** information
- **Acronyms**
- Established **dates**
- Additional **identifier**
  - Wikipedia/Wikidata IDs
  - ISNI
  - ROR IDs
  - HESA



Possible affiliation resolution workflow for one-off clean-up or continuous enrichment



Dimensions

Part of DIGITAL science

Organization disambiguation - API example

Dimensions

Part of DIGITAL science

See you in 10!

Dimensions

Part of DIGITAL science

Dimensions

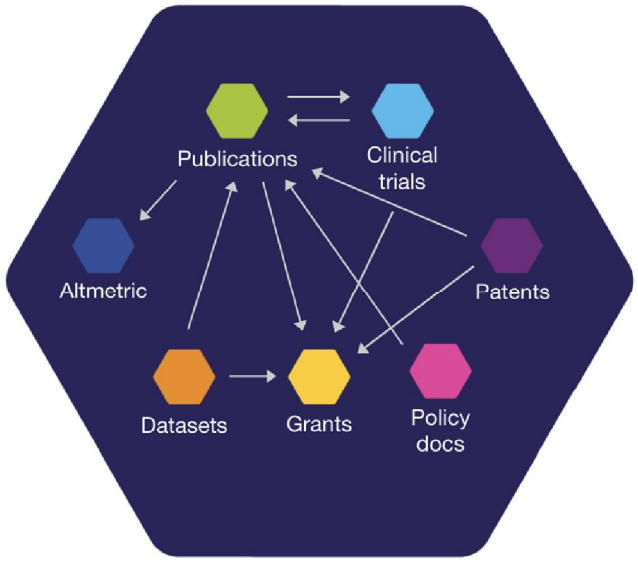
Classify content

How Dimensions can help to classify records

Part of DIGITAL science

How we approach classifications in Dimensions

- Journal level classifications not sufficient for the broader approach (grants, patents, trials etc.)
- Instead, we are using the content/text to categorise documents based on their ‘substance’ rather than only one metadata element
- The categorisation happens via NLP and machine learning based classification schemes
- Based on this automated and scalable approach we are able to integrate 1-n classifications
- And the classify functionality is available for the user to categorise documents outside of Dimensions

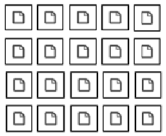


Standing on expert's shoulders - getting to a training set

Starting point

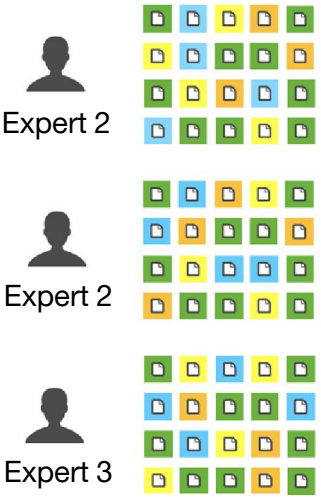
FOR classificatio n system

&

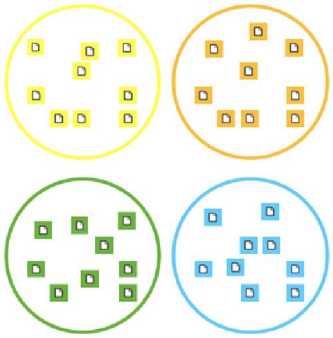


Document set without categories

FOR codes assigned in ERA process by experts and researchers



Result: categorised docs - ready to use as a training set

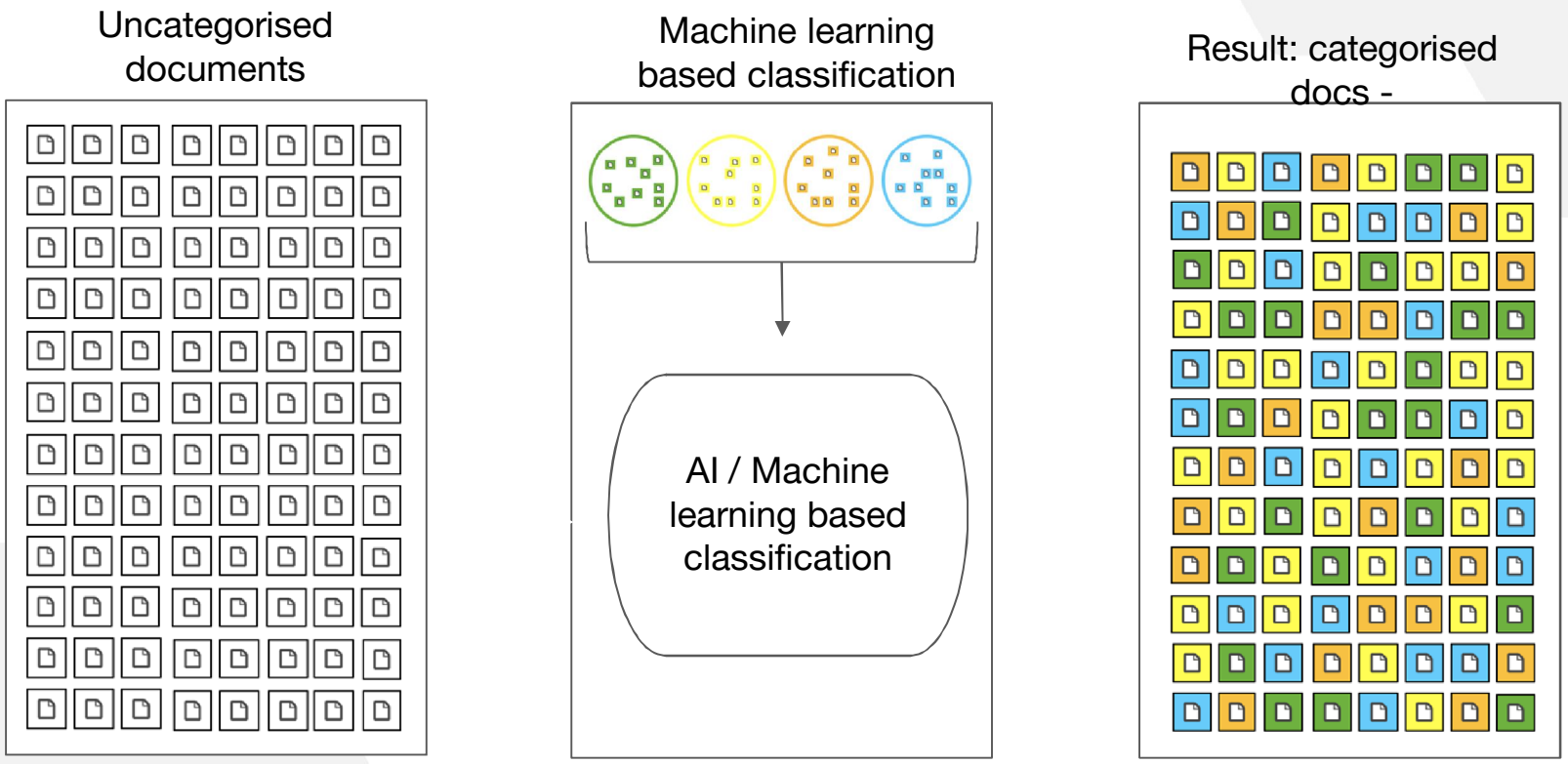


Category Definitions Generated

Dimensions

Part of DIGITAL science

With this model we can categorise millions of documents in minutes...



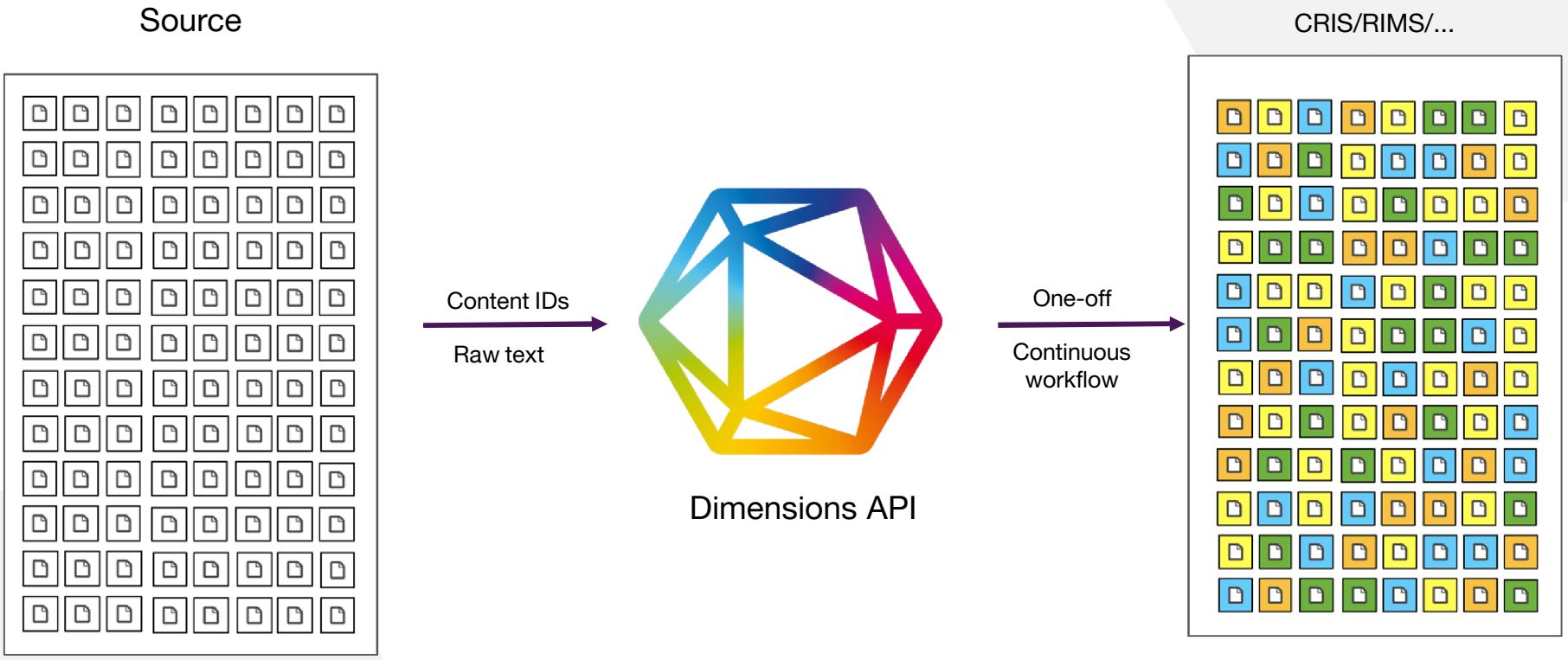
Dimensions

Part of DIGITAL science

Classifications in Dimensions - selection of main systems

Area	Classification	From	Granularity
All research areas	**Fields of Research - FOR codes**	Australia/New Zealand, used in national assessment exercise	176 classification options
All research areas	**Units of assessment - UoAs**	UK, used in REF exercise	34 classification options
Special interest	**Sustainable Development Goals (SDGs)**	United Nations - developed to put a massive focus on challenges human kind (still) faces	17 classification options
Domain specific	**Research, Condition, and Disease Categorization (RCDC)**	The NIH uses since 2005 the RCDC classification to report on their funding activities	295 classification options
Domain specific	**Health Research Classification System Health Categories (HRCS_HC and HRCS_RAC)**	Two Health related classification with health/disease area and research activity categories, used by UK based funders, development led by MRC	77 classification options
Domain specific	**ICRP Common Scientific Outline (ICRP_CSO and ICRP_CT)**	Cancer specific classification developed by ICRP, used by more than 120 funders globally to align	102 classification options

A sample application workflow for automated classification



Dimensions

Part of **DIGITAL** science

Dimensions API

View API documentation

The Dimensions API provides access to Dimensions data directly, and makes it possible to retrieve results to precise and complex queries. These are performed using the Dimensions Search Language (DSL), our own domain specific language. DSL expresses queries using terms and structures relevant to the Dimensions data. Explore real-world applications of the API at the Dimensions API Lab ↗, an open-source repository of Jupyter ↗ notebooks demonstrating how to carry out common scholarly analytics tasks.

Query

Type a query

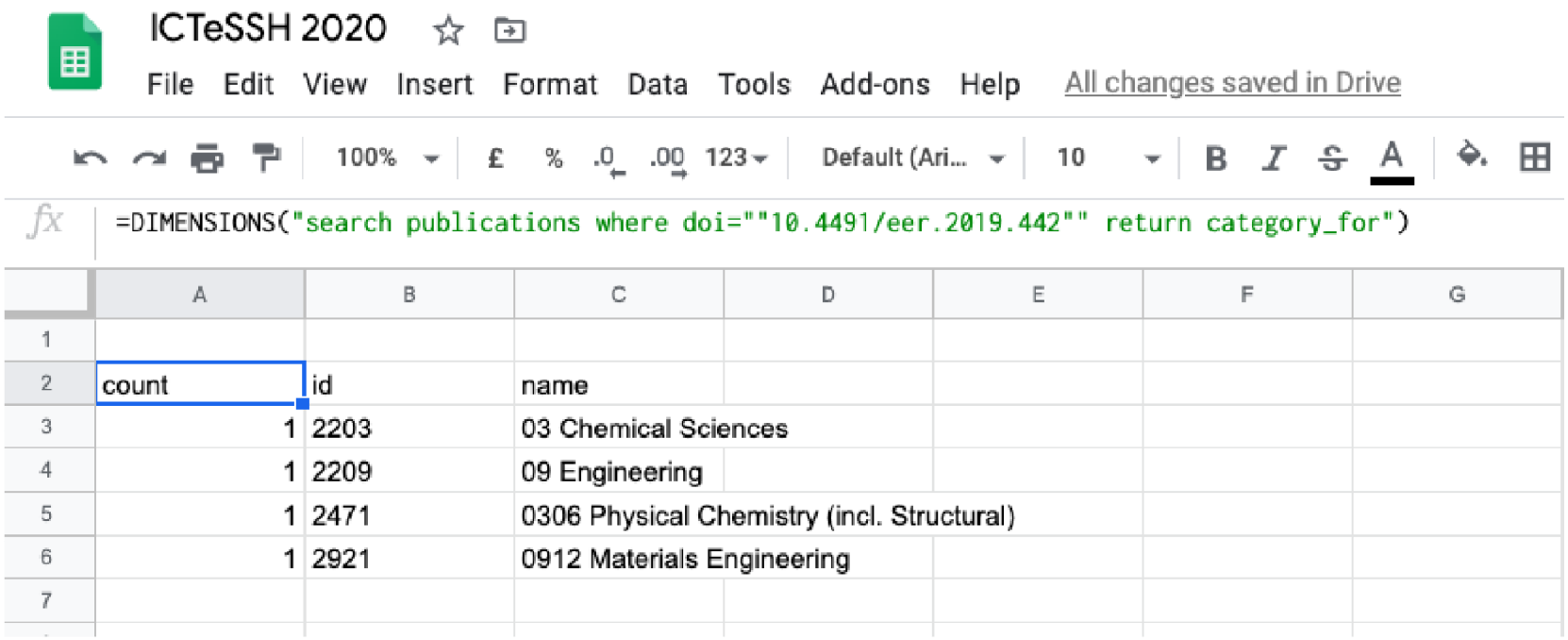
```
classify(
    title="Burnout and intentions to quit the practice among community pediatricians: associations with specific professional activities",
    abstract="BACKGROUND: Burnout is an occupational disease expressed by loss of mental and physical energy due to prolonged and unsuccessful coping with stressors at work. A prior survey among Israeli pediatricians published in 2006 found a correlation between burnout and job structure match, defined as the match between engagement with, and satisfaction from, specific professional activities. The aims of the present study were to characterize the current levels of burnout and its correlates among community pediatricians, to identify changes over time since the prior survey, and to identify professional activities that may reduce burnout. METHODS: A questionnaire was distributed among pediatricians both at a medical conference and by a web-based survey. RESULTS: Of the 518 pediatricians approached, 238 (46%) responded to the questionnaire. High burnout levels were identified in 33% (95% CI:27-39%) of the respondents. Higher burnout prevalence was found among pediatricians who were not board-certified, salaried, younger, and working long hours. The greater the discrepancy between the engagement of the pediatrician and the satisfaction felt in the measured professional activities, the greater was the burnout level (p < 0.01). The following activities were especially associated with burnout: administrative work (frequent engagement, disliked duty) and research and teaching (infrequent engagement, satisfying activities). A comparison of the engagement-satisfaction match between 2006 and 2017 showed that the discrepancy had increased significantly in research (p < 0.001), student tutoring (P < 0.001), continuing medical education and participation in professional conferences (P = 0.0074), management (p = 0.043) and community health promotion (P = 0.006). A significant correlation was found between burnout and thoughts of quitting pediatrics or medicine (p < 0.001). CONCLUSIONS: Healthcare managers should encourage diversification of the pediatrician's job by enabling greater engagement in the identified anti-burnout professional activities, such as: participation in professional consultations, management, tutoring students and conducting research.",
    system="FOR")
```

Run

Results

Copy to clipboard


```
▼ {  1 item
  ▼ "FOR" : [  1 item
    ▼ 0 : {  2 items
        "id" : "3177"
        "name" : "1117 Public Health and Health Services"
      }
  ]
}
```


ICTeSSH 2020
File Edit View Insert Format Data Tools Add-ons Help
All changes saved in Drive
100%
Default (Ari...
10
=DIMENSIONS("search publications where doi=""10.4491/eer.2019.442"" return category_for")
count
id
name
1 2203 03 Chemical Sciences
1 2209 09 Engineering
1 2471 0306 Physical Chemistry (incl. Structural)
1 2921 0912 Materials Engineering

Dimensions

Extract concepts

How keywords can be generated automatically

Part of DIGITAL science

Automatically assigned keywords for better Discoverability

It is generally recommended to exercise caution in applying trauma-focused treatment to individuals with posttraumatic stress disorder (PTSD) and comorbid borderline personality disorder (BPD). Objective: To investigate the effects of a brief, intensive, direct trauma-focused treatment programme for individuals with PTSD on BPD symptom severity. Methods: Individuals (n = 72) with severe PTSD (87.5% had one or more comorbidities; 52.8% fulfilled the criteria for the dissociative subtype of PTSD) due to multiple traumas (e.g. 90.3% sexual abuse) participated in an intensive eight-day trauma-focused treatment programme consisting of eye movement desensitization and reprocessing (EMDR) and prolonged exposure (PE) therapy, physical activity, and psychoeducation.

Dimensions API

Extract concepts:

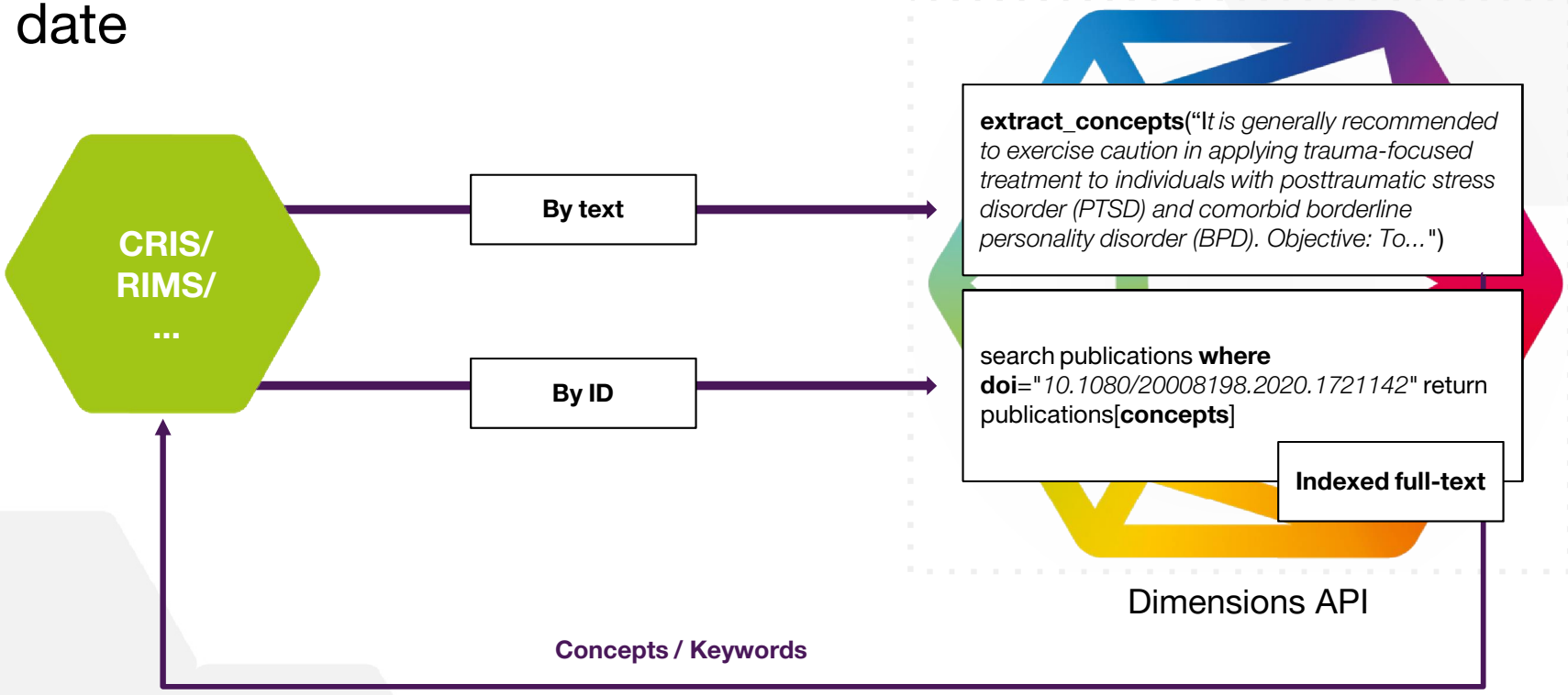
- Generating free keywords
- Domain independent
- Unlimited vocabulary
- Supports emerging terms

- Trauma
- Disorders
- PTSD
- treatment programs
- borderline personality disorder
- personality disorders
- severe PTSD
- individuals
- exposure therapy
- eye movement desensitization
- psychoeducation
- stress disorders
- symptom severity
- multiple trauma
- physical activity
- BPD symptom severity
- severity
- therapy
- desensitization
- treatment
- day trauma
- caution
- program
- activity
- effect

Dimensions

Part of **DIGITAL** science

One-off or part of continuous workflow:
Make sure keywords are relevant, always added and up to date



Analysis ideas

	Azure	Google Cloud	Amazon EC2	AWS	security	serverless computing	fog computing
Azure	106	19	6	7	23	1	3
Google Cloud	19	50	2	5	7	1	0
Amazon EC2	6	2	93	7	16	0	3
AWS	7	5	7	107	20	17	1
security	23	7	16	20	3680	2	273
serverless computing	1	1	0	17	2	45	1
fog computing	3	0	3	1	273	1	1204

Analysis ideas

Concepts to first appear in 2019	Frequency
Trans-Omics	12
precision medicine programs	10
dermal fibroblasts	9
induced pluripotent stem cell line	8
NHLBI Trans-Omics	7
aging-related diseases	7
pediatric cancer	7
fungal isolates	6
independent genome-wide significant loci	6
nucleotide identity	6
phenotypic divergence	6

For subset of data in FoR 0604 Genetics

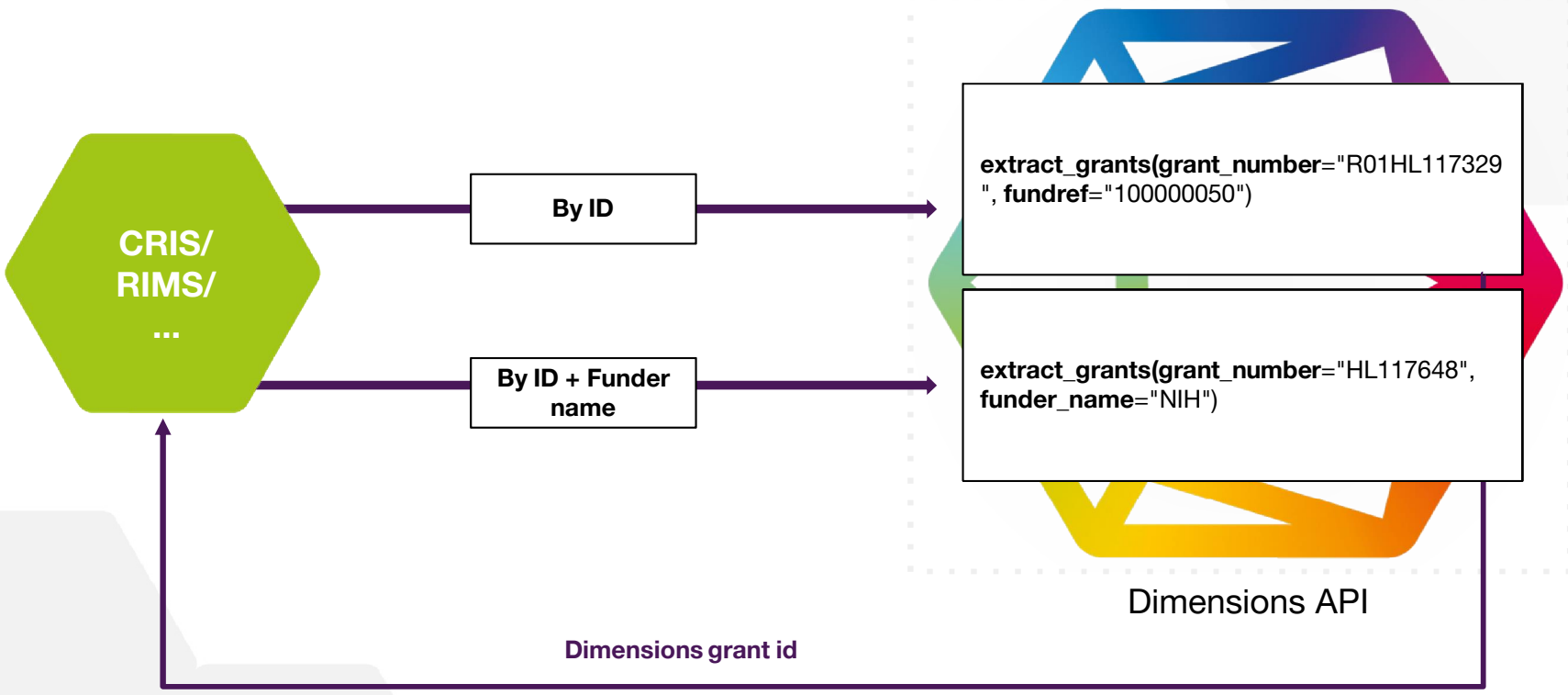
Dimensions

Extract grants

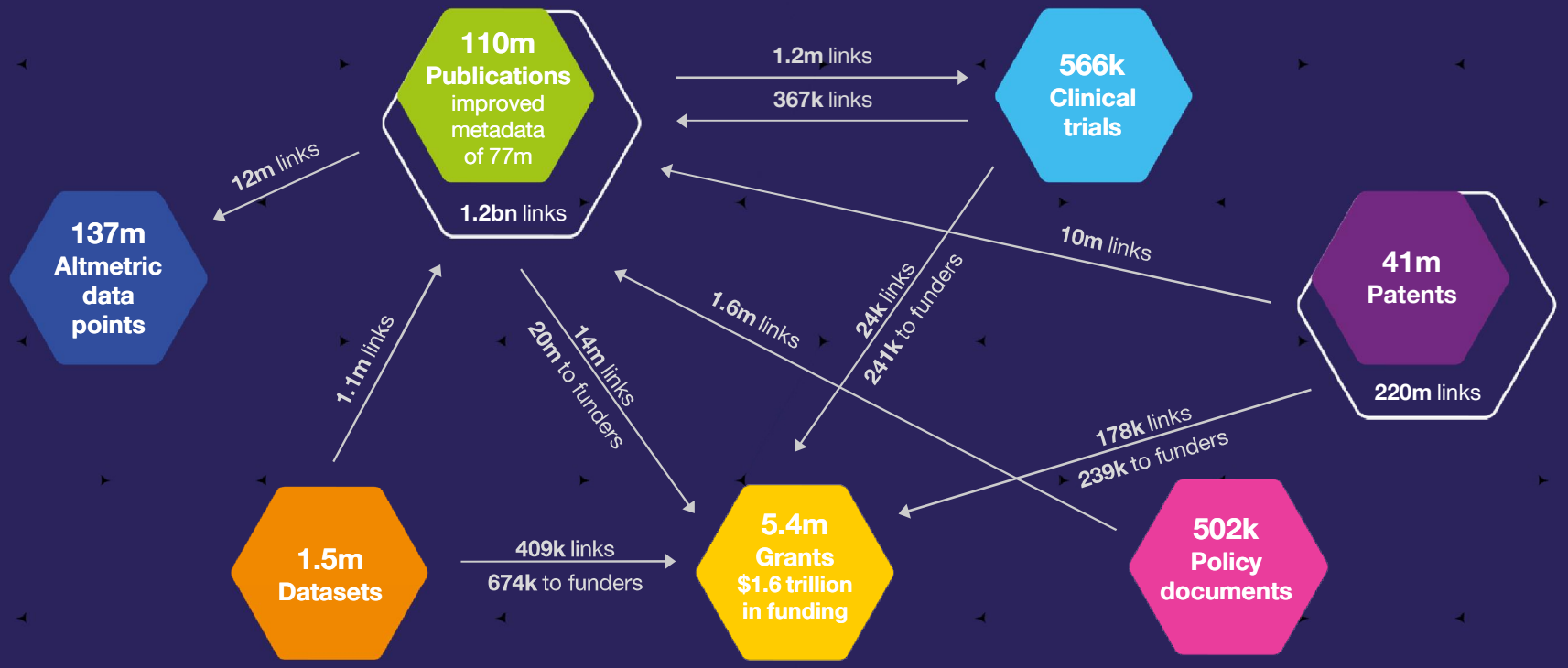
How grant ids can be extracted automatically

Part of DIGITAL science

One-off or part of continuous workflow: Connect the data



Dimensions data world



Dimensions

Status: June 2020

Part of **DIGITAL** science

Q&A

Dimensions

Part of DIGITAL science

Thank you!

Cristina Huidiu

c.huidiu@digital-science.com

Dimensions

Part of DIGITAL science