Analyzing a CRIS: From data to insight in university research

euroCRIS
Current Research Information Systems

SIGMA
helping universities succeed

“The amount of data we produce every day is truly mind boggling”.

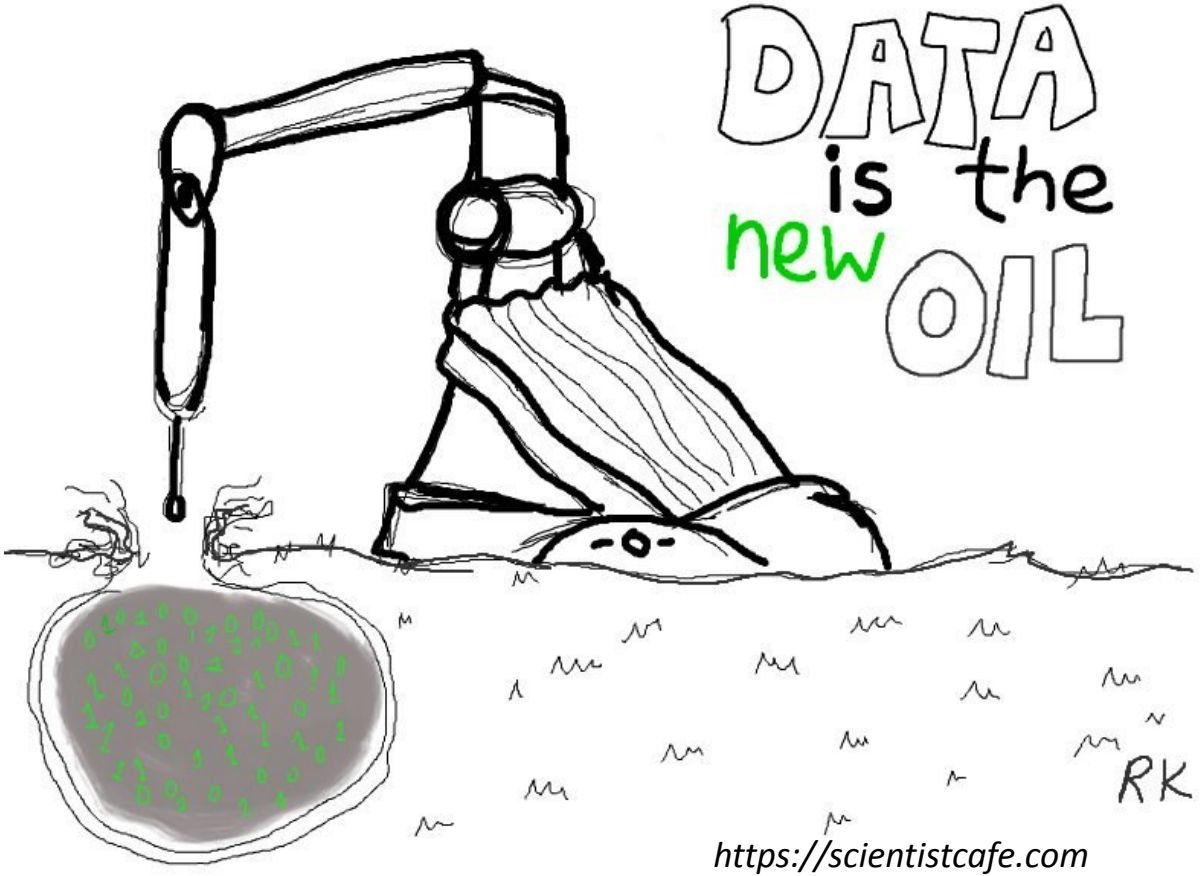
Data is growing faster than even before and by the year 2020, about 1.7 Mb of new information will be created every second for every human being on the planet.

Bernard Marr (Forbes 2018)

The famous Internet minute…

2018 What happens in an Internet minute

Is so relevant, that we can say:



https://scientistcafe.com

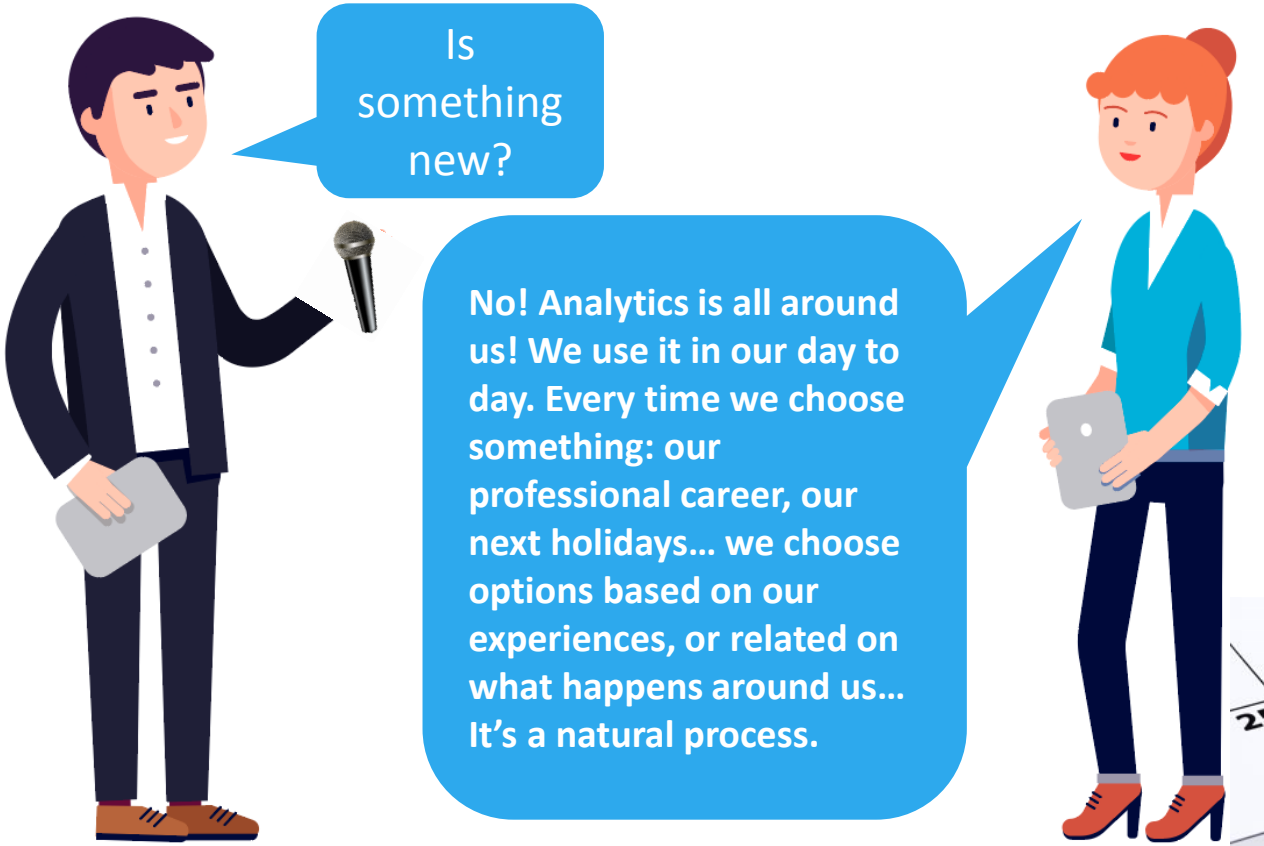
So much data….

Don't know what to do!!

The human brain cannot process so much data…

And that's why we start talking about **data analytics….**

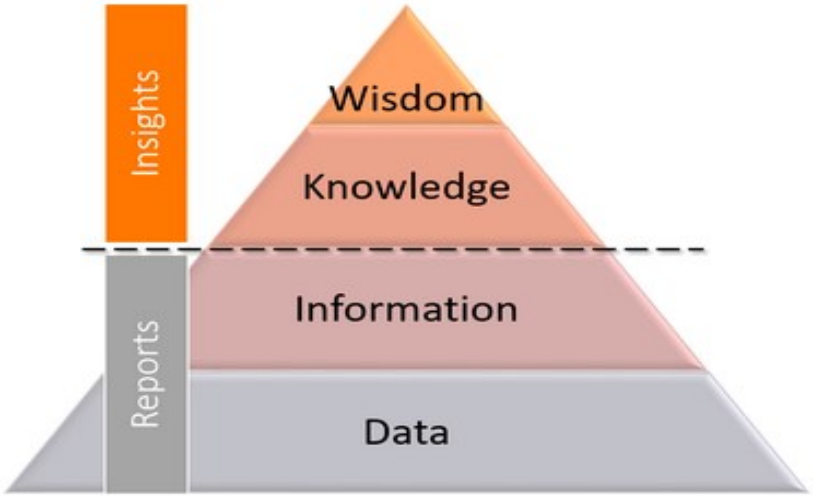
Let's talk about **data analytics**… an interview:



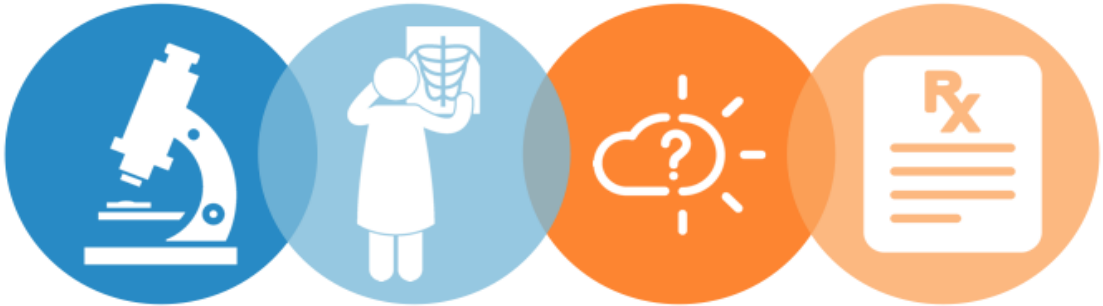
Let's talk about **data analytics**… an interview:

So we can define:

“Data analytics is the ability to collect and use data to generate insights that inform fact-based decision making.”

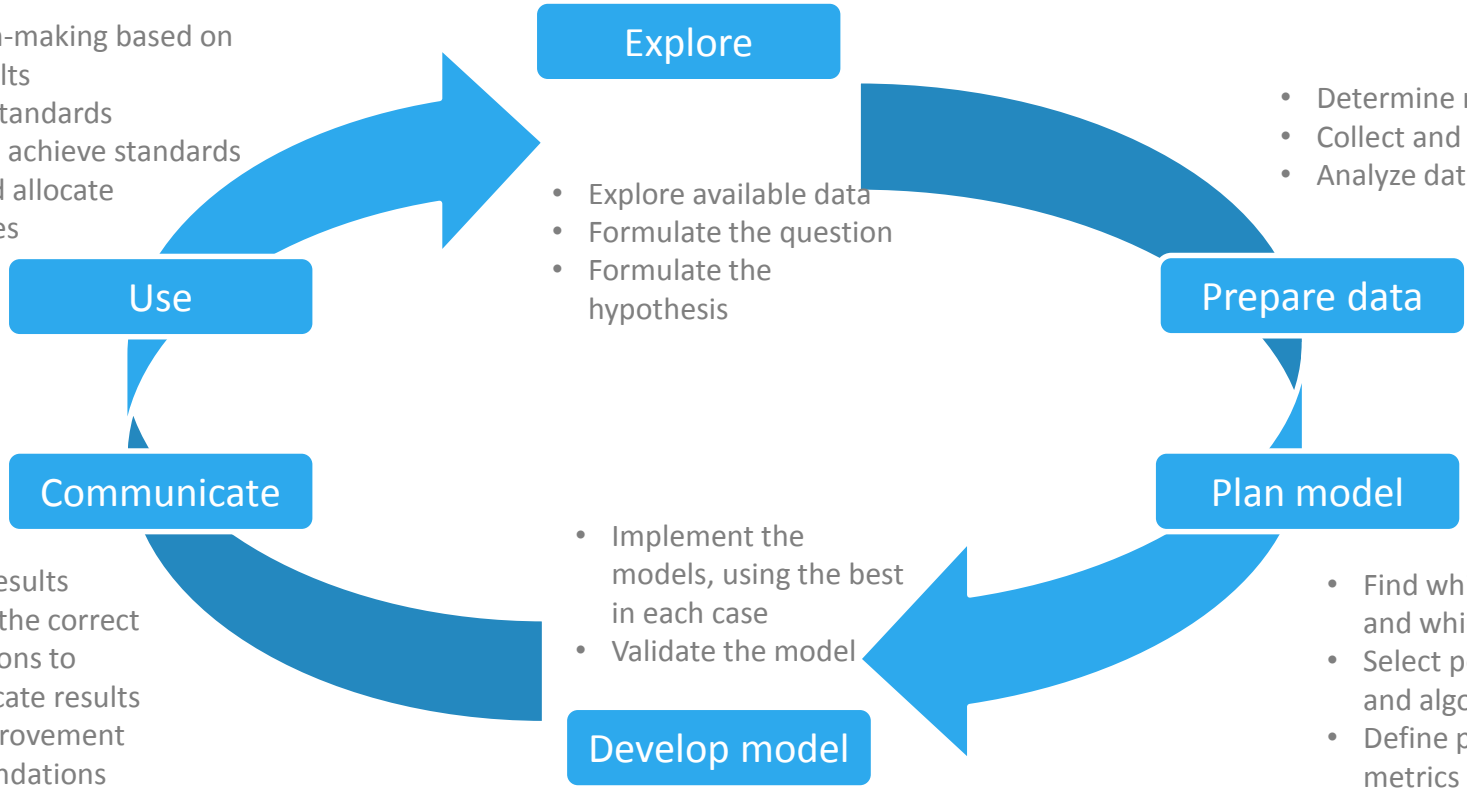


We can classify the data analytics process in four stages:



Its easy to conclude that more value is added as you advance in the list below!

Data analytics lifecycle is an iterative process:



- Explore available data
- Formulate the question
- Formulate the hypothesis

- Determine necessary data
- Collect and clean data
- Analyze data consistency

- Find which variables tells and which predicts
- Select possible models and algorithms to use
- Define performance metrics

- Implement the models, using the best in each case
- Validate the model

- Perform results
- Generate the correct visualizations to communicate results
- Make improvement recommendations

- Decision-making based on the results
- Define standards
- Goals to achieve standards
- Plan and allocate resources

So It's not trivial to find results through data, it is necessary to draw an elaborate plan and take some actions. New kind of experts appeared with new skills…

Data Scientist

also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:
SQL, Python, R

Data Engineers

also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:
Hadoop, NoSQL, and Python

Data Analysts

also known as business Analysts.

They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

Will use programmes such as:
Excel, Tableau, SQL

Source: https://medium.com/@vegi/data-scientist-vs-data-analyst-vs-data-engineer-using-word-cloud-902ab83d0879

Having all this in mind, at SIGMA, since 2011, we are developing analytical environments to give universities tools to easy their information analysis both in research and academic areas.

In this way, we have experienced **data scientists & data engineers** that create analytical environments for Teaching, Students, Economic management, Research and so on.

Thanks to our work close to the universities, we also have **data analysts** profiles experienced in higher education processes (both academic and research)

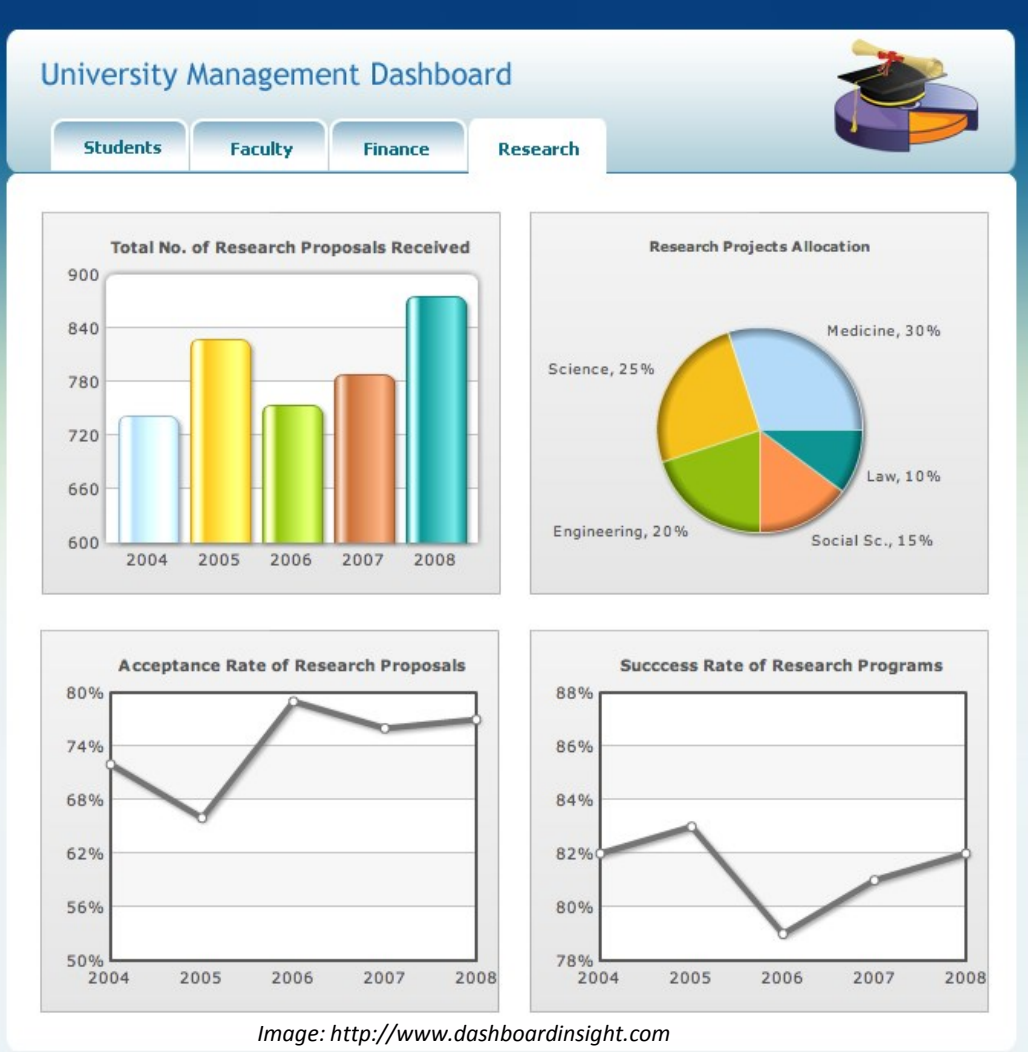


Image: http://www.dashboardinsight.com

Focusing in research area, it is easy to do when we have the enough maturity in the CRIS system that incorporates the largest number of scientific information of the university with quality.

To develop a project of data analytics and **focusing on the Research area**, we followed the next **6 steps**:

1

Classify the Research scope in distinctive knowledge areas

2

Translate the CRIS (CERIF compliant) model to an analytical data model

3

Define main indicators in each knowledge area with university experts collaboration

4

Implement (calculate if necessary) the indicators and main analytical outputs

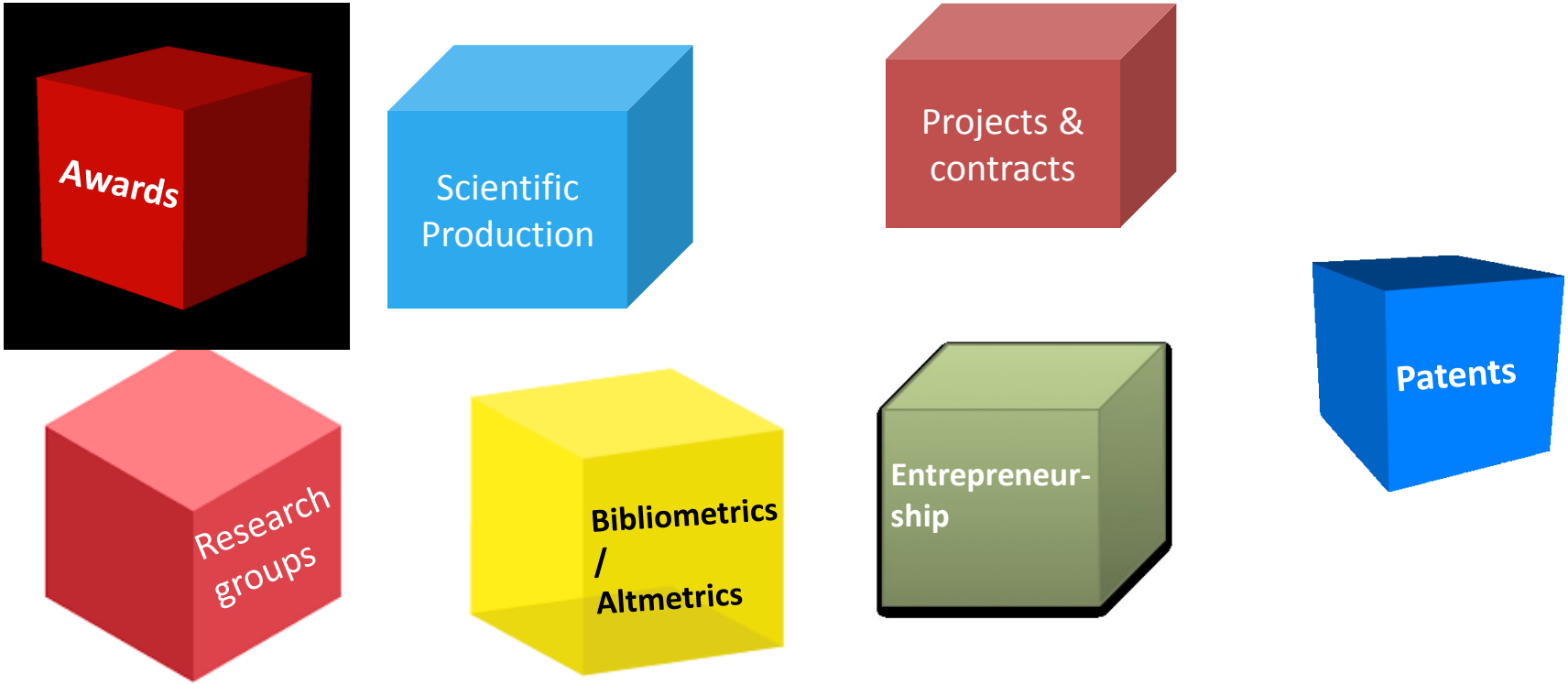
5

Validate the data quality (data cleansing)

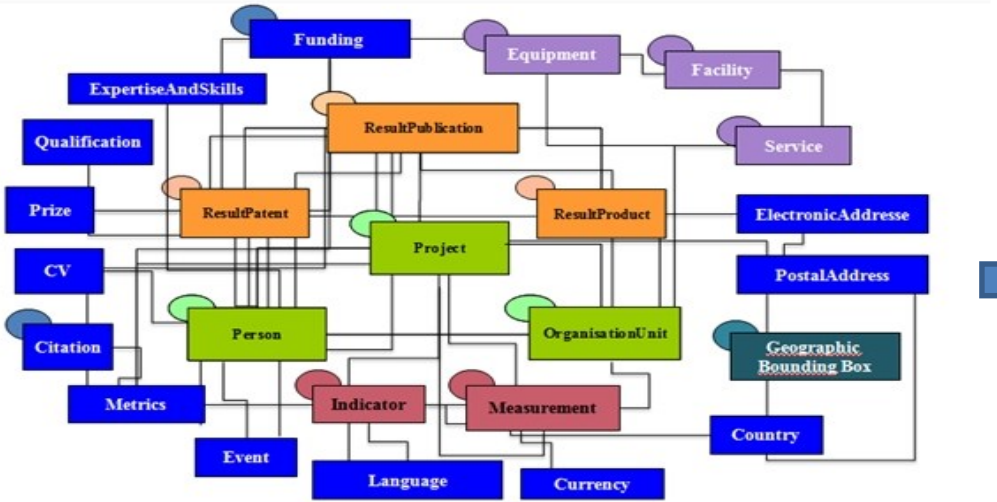
6

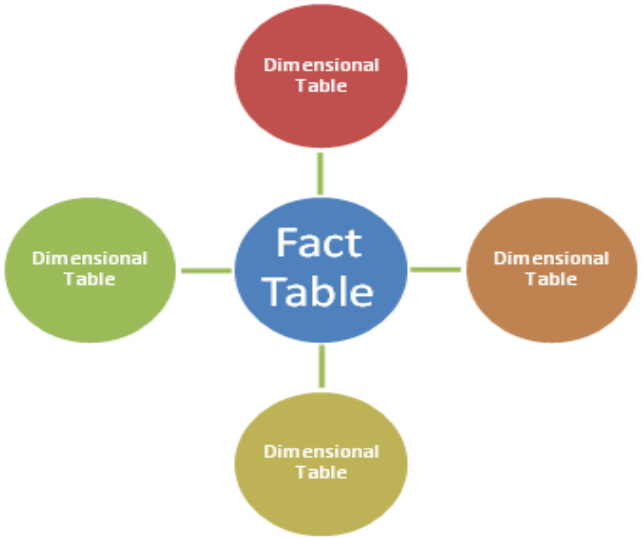
Open product on the university, improve process (iterative)

1. Classify the Research scope in distinctive knowledge areas, called analytical cubes:



2. Translate de CRIS (CERIF) model into a star model, unlinking the analytical environment from the transactional environment.





3) Define the main indicators in each knowledge area or cubes with the university experts collaboration. i.e:

Scientific Production

We can analyze many facts:

- # of citations at different levels
- # of publications of all kinds
- Reputation (author&publication)
- Rankings of different kinds
- ….

→

Through many dimensions:

- Author data
- Year of publication
- Impact (wos, scopus, google scholar…)
- Citations (wos, scopus, google scholar…)
- SJR
- ….

Projects & Contracts:

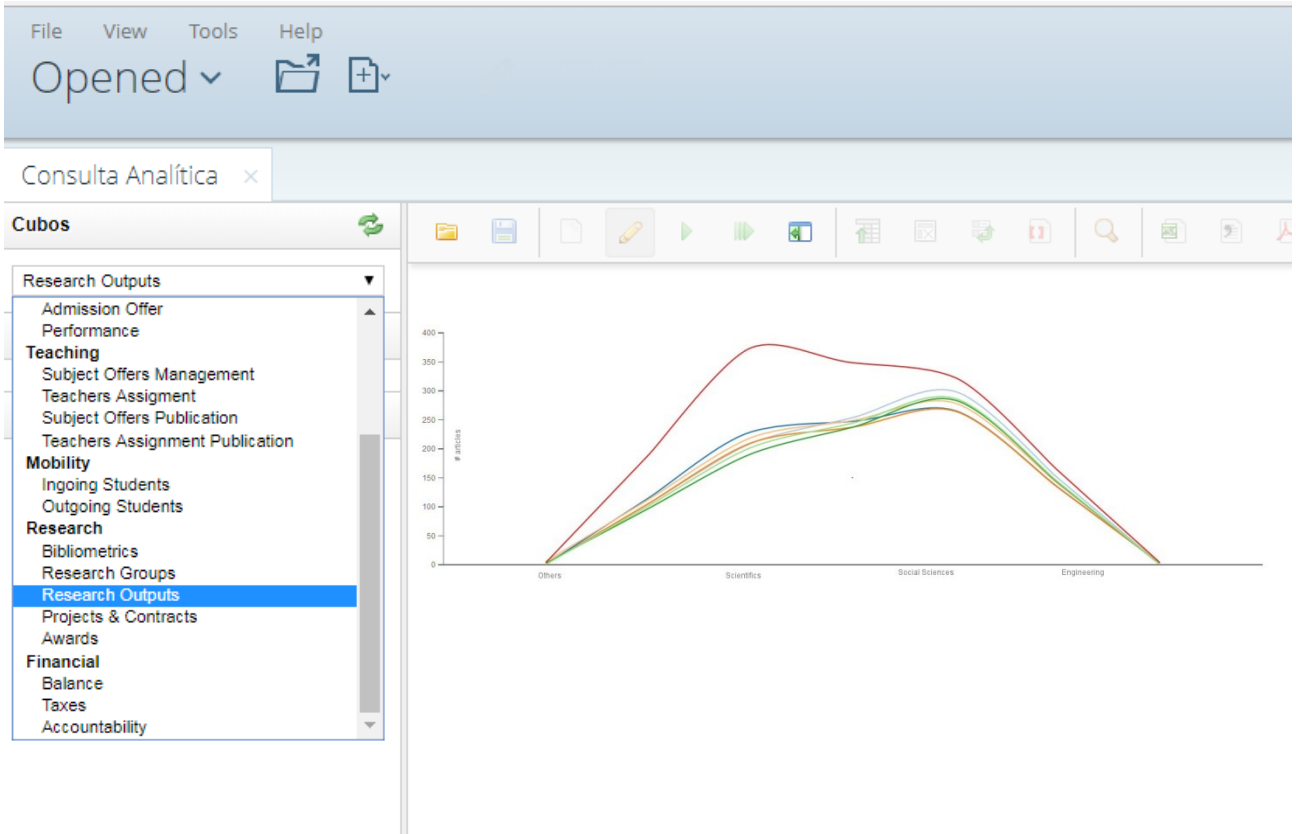
We can analyze many facts:

- # of projects
- # of active projects
- Expected income
- Real income
- Total income
- % of public financing
- % of private financing
- ….

Through many dimensions:

- Award
- Grant
- Associates
- Country
- ….

4. Implement (calculate if necessary) the indicators and main analytical outputs with an analytical tool

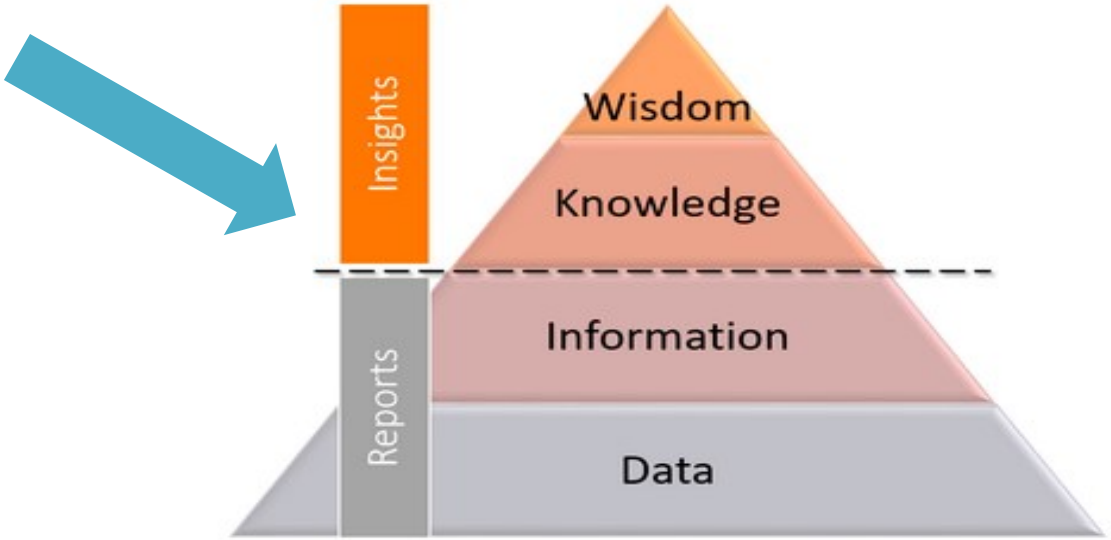


5. Validate the data quality (avoiding redundancies, referential integrity, range of values…)

- Data deduplication
- Quality check
- Data normalization
- Data standardization

6. Open product on the university, improve the process (if necessary) → start the iterative process

When we have knowledge, we have the first insights and we can act based on it…



Examples: we can analyze all the authors of the institution….

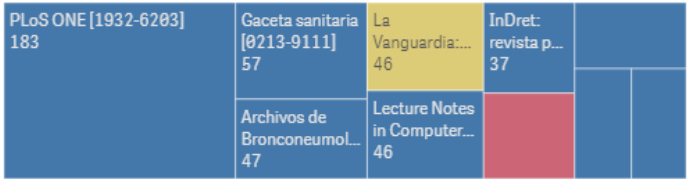
(dashboards)

Publications & Bibliometrics

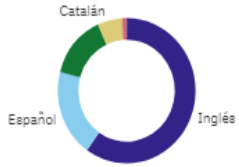
Year

Num. of Articles	Num. of Authors	Num. of Journals	Scopus Citations	WOS Citations
Num. Articles	Num. Authors	Num. Journals	Total Scopus Citations	Total WOS Citations
5,5k	1,49k	2,28k	61,76k	57,11k

Top 10 Journals by Articles



Articles by Language



Authors & Articles by Genre

Top Citations Articles

Article	Total Scopus Citations
'If you are good, I get better': the role of social hierarchy in perceptual decision-making	15
'Woe Betide Us If They Win!': National Socialist Treatment of the Spanish 'Volunteer' Workers	1
"20209C-T" a variant mutation of prothrombin gene mutation in a patient with recurrent pregnancy loss	0
"A Casual Delight" (Mary Stuart Boyd)	0
"A City Alive" (Madge Macbeth)	0
"A Dream of Bulls" (Charles Francis Schroeder)	0

Top H-Index Authors

Authors	Max H-Index Scopus	Max H-Index WOS
SERRA GUIGO, RODERIC	65	49
MARTINEZ COMAS, DAVID	43	46
DEU SUNYER, JORDI	72	45
LOPEZ MALDONADO, RAFAEL	53	44
BOQUÉ ANTÓ, JOSE MARIA	2	43
BUSQUETS BERTRANPETIT, JAUME	52	40
MAJO CALAFELL, FRANCESC D'ASSIS	48	39
GARRIGA POSAS, FRANCESC	36	38

Examples: …or only one author

Publications & Bibliometrics

Year

Num. of Articles

Num. Articles

Num. of Authors

Num. Authors
1

Num. of Journals

Num. Journals
22

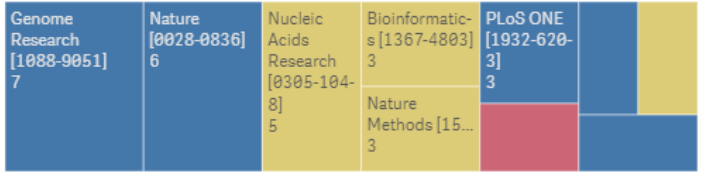
Scopus Citations

Total Scopus Citations
7,51k

WOS Citations

Total WOS Citations
7,31k

Top 10 Journals by Articles



Articles by Language

Authors & Articles by Genre

Top Citations Articles

Article	Total Scopus Citations
Absence of canonical marks of active chromatin in developmentally regulated genes	4
An encyclopedia of mouse DNA elements (Mouse ENCODE)	132
An integrated encyclopedia of DNA elements in the human genome	3022
ASPic-GeneID: a lightweight pipeline for gene prediction and alternative isoforms detection	0
Assessment of transcript reconstruction methods for RNA-seq	90
BLUEPRINT to decode the epigenetic signature written in blood	0

Top H-Index Authors

Authors	Max H-Index Scopus	Max H-Index WOS
SERRA GUIGO, RODERIC	65	49

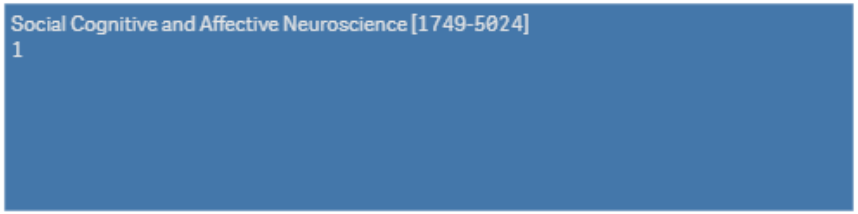
Examples: …or the authors from one article

Publications & Bibliometrics

Year

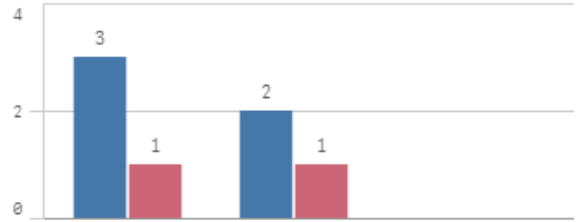
Num. of Articles	Num. of Authors	Num. of Journals	Scopus Citations
Num. Articles	Num. Authors	Num. Journals	Total Scopus Citation:
1	5	1	15

Top 10 Journals by Articles



Articles by Language

Authors & Articles by Genre



Top Citations Articles

Article	Total Scopus Citations
'If you are good, I get better': the role of social hierarchy in perceptual decision-making	15

Top H-Index Authors

Authors
GALLES SEBASTIAN, NURIA
. DECO, GUSTAVO
. PANNUNZI, MARIO
GARCIA SANTAMARIA, HERNANDO
GIMENO AYNETO, ALBA

Examples: …we can elaborate university author rankings based on the main impact indicators that we have stored in the CRIS through the years

Rankings

Year

Top H-Index Authors

Authors	Total Scopus Citations	Total WOS Citations	Max H-Index Scopus	Max H-Index WOS
Totales	**61762**	**57113**	**-**	**-**
SERRA GUIGO, RODERIC	7509	7314	65	49
MARTINEZ COMAS, DAVID	523	515	43	46
DEU SUNYER, JORDI	761	755	72	45
LOPEZ MALDONADO, RAFAEL	729	663	53	44
BOQUÉ ANTÓ, JOSE MARIA	424	425	2	43
BUSQUETS BERTRANPETIT, JAUME	508	468	52	40
MAJO CALAFELL, FRANCESC D'ASSIS	49	41	48	39
GARRIGA POSAS, FRANCESC	191	161	36	38
MADUEÑO GARCIA DE HERREROS, ANTONIO	349	348	43	36
GRANADOS VALVERDE, OLGA	199	181	33	28
PUJADAS SERRA, CONSUELO	306	306	33	28
GARRETA GALI, JORDI	21	15	25	28
DE ROVIRA BENACH, JUAN	164	143	33	27
TRULL LLORETA, JOSÉ	486	386	33	27
AYMERICH GARCIA, JUDITH	716	649	32	25
BIGAS LOPEZ, NURIA	635	633	-	24

Examples: … we can geolocate the different congresses where the university researches take part the most.

Congress

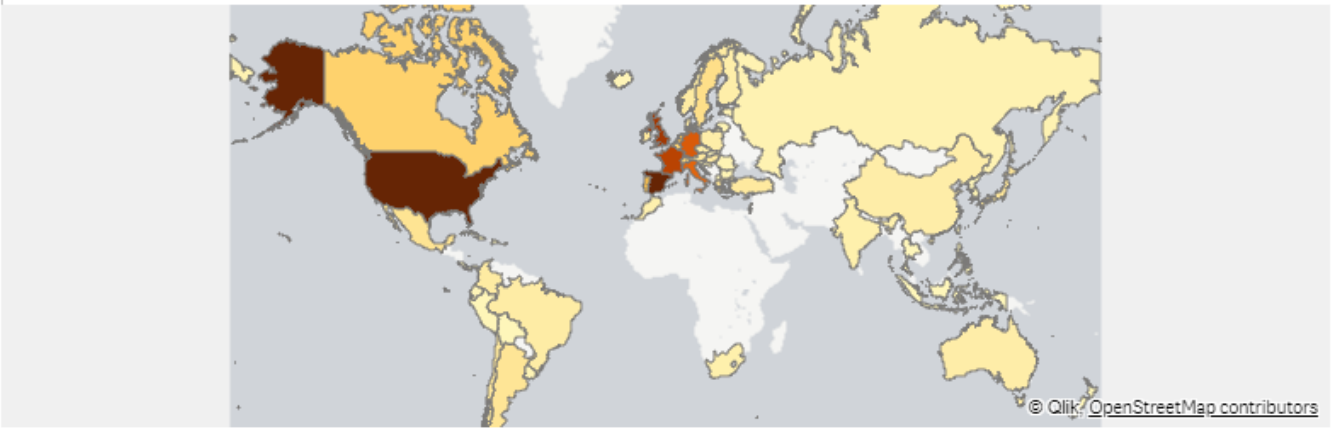
Year

Congress distribution by country

Congress country
Capa de área

0 200

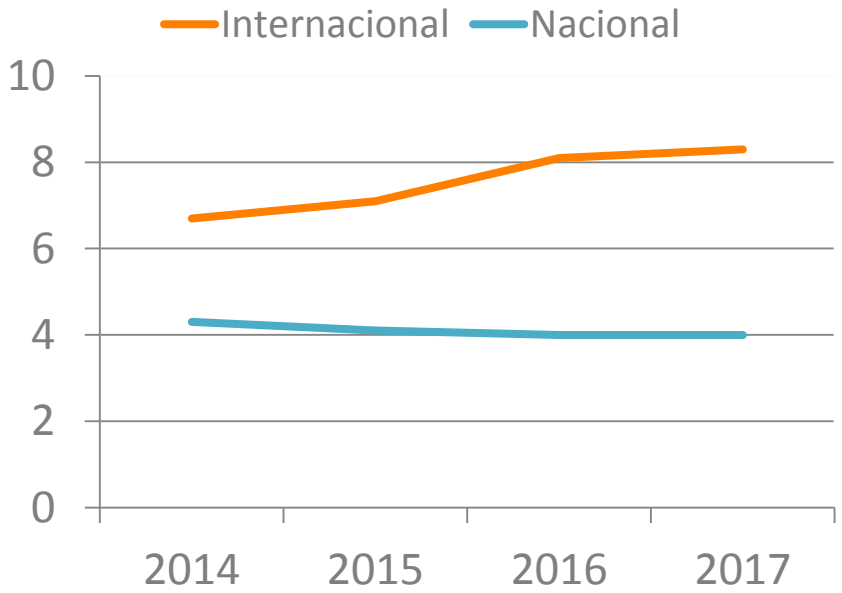
Total Congress



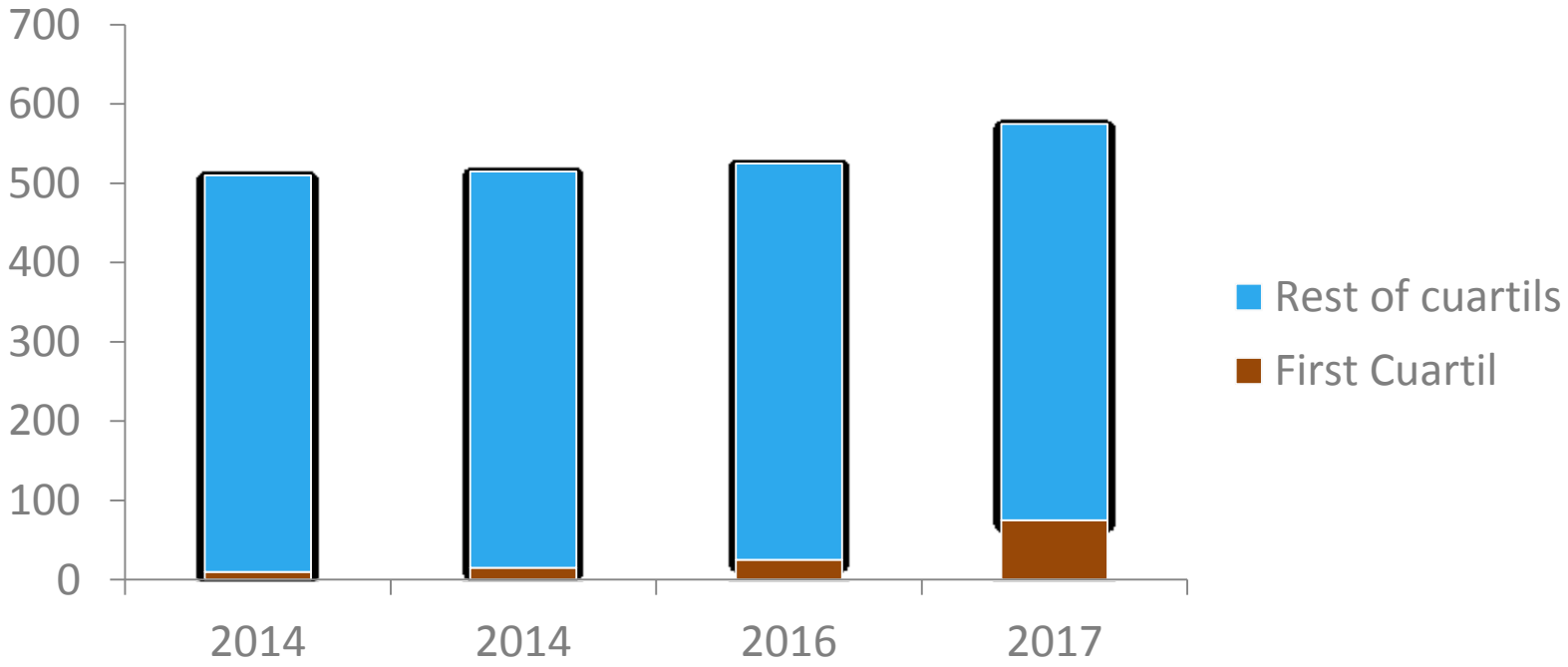
Congresses

Congress	Congre
Workshop on Collateral Sanctions	Estado
Workshop on Criminal Records	Estado
Workshop on Databases and Corpora in Linguistics	Estado
Workshop pn Forensic Authorship Identification	Estado
XVIII Biennial International Conference on Infant Studies	Estado
"Change in political attitudes:Panels and experiments" Workshop	España
"The Friars and their Influence in Medieval Spain¿	España
+ Democracia: II Congreso de Análisis Político Crítico	España
1a Jornadas transfronterera de la recerca Empordà/Rosselló (s. VI-I a.n.e.).	España
1a Jornades Transfrontereres de la recerca arqueològica Empordà/Rosselló (s. VI-I a.n.e.)	España

Examples: … We can show the co-authoring evolution through the years and thus we could discover if the internalization degree of the researchers of the institution advance correctly…



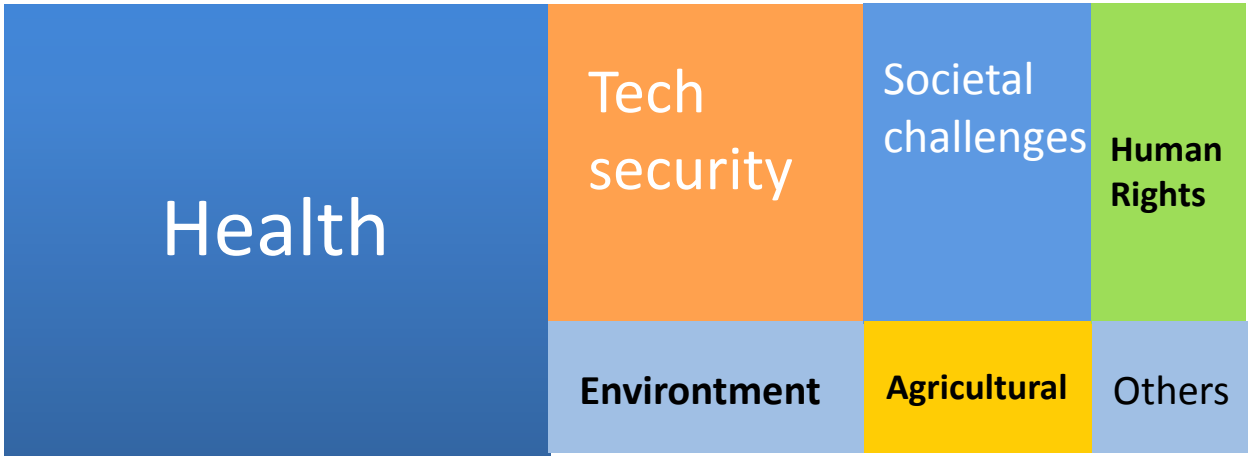
Examples: …we can know the evolution of authors of my institution that publish in first quartile journals



Examples: …and as a researcher, the state of my research group. For example, the new projects of the group through the years, to determine the group evolution.

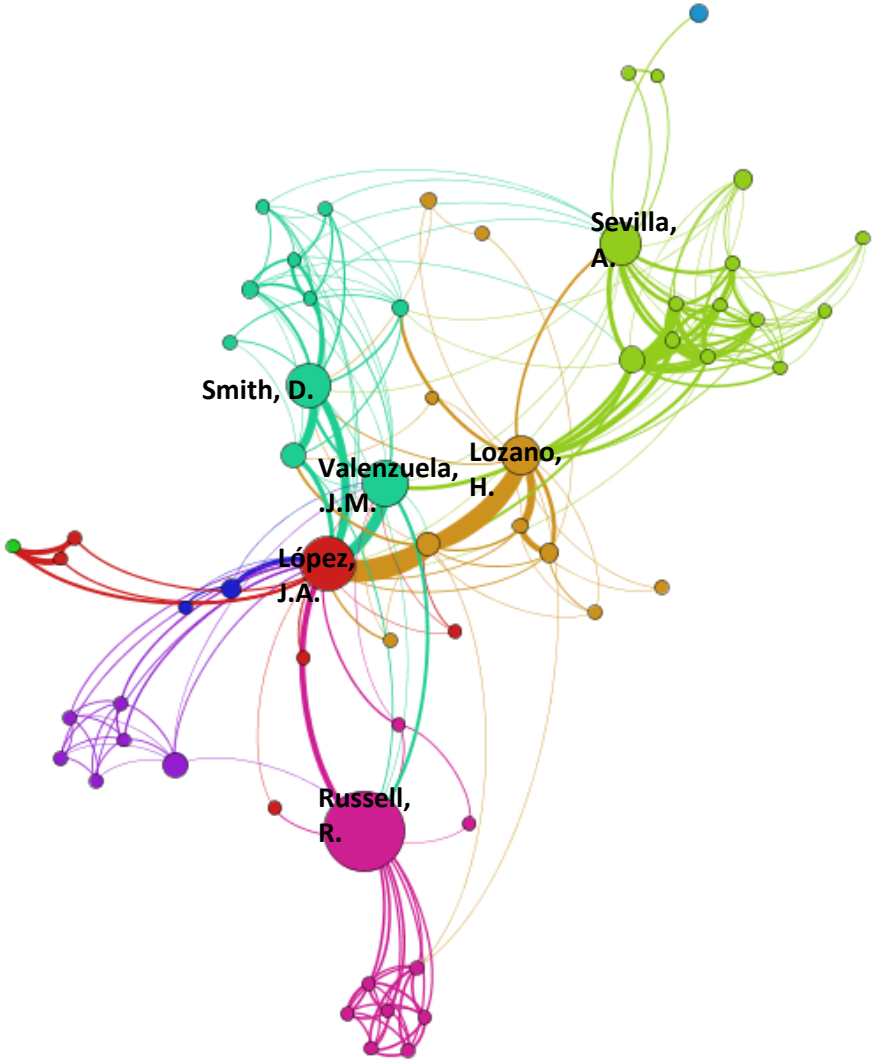
	Year										
Research Group Title	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	
214 - Propiedades Nanométricas de la Materia	8	8	3	6	1	2	0	6	2	0	
321 - G-FOR (Grupo de Fotónica, Información Cuántica y Radiación y Dispersión de Ondas)	9	5	6	7	0	2	0	2	3	0	

Examples… we can track the projects discovering which researchers have more funding and in which areas, ordering by the project results impact, determining in which areas is the university, top. This can provide an evidence of where is best the university and could focus on it.



Examples… Finally, we can discover the collaboration degree between the researchers of our institution, through the researchers interactions.

We can discard the weak interactions and which researchers have more influence.



To have into account:

1) We need the **time** variable that shows the evolution, to explain what had happened
2) The information should be from **more than one institution**, providing comparatives between institutions

With all of this information we can know exactly which are the impact authors of our institution and if they are always the same every year or not.
Also a researcher could track the impact evolution of his research group

Conclusions:

We have the **descriptive and diagnostic states** complete

- ❑ The research information is now FAIR in order to ease the reporting, analyses and access to the information
- ❑ Users can see what happens in the past and why it happens
- ❑ The process is continuously improved
- ❑ This is more a **service** than a product

We are now boarding the **predictive state** in order to find the facts (using data mining and machine learning technologies) to see what could happen in the future, based on what had happened

Conclusions:

We should not be afraid to face a project like this if we have the enough data maturity in the CRIS (CERIF compliant)

(data with quality, accessible and reusable enough)

Thank you so much for your attention.

SIGMA

helping universities succeed

Anna Guillaumet

@annaguillaumet

anna.guillaumet@sigmaaie.org