Corpus Viewer platform relies on Natural Language Processing (NLP), Machine Learning (ML) and Machine Translation (MT) to analyze structured metadata and unstructured textual data in large document corpora. The platform allows the decision maker and the policy implementer the possibility of analyze R&D&i information space (mainly patents, scientific publications and public aids) for evidence and knowledge-based policy making and implementation. It relies, among other techniques, on topic modeling and graph analysis.

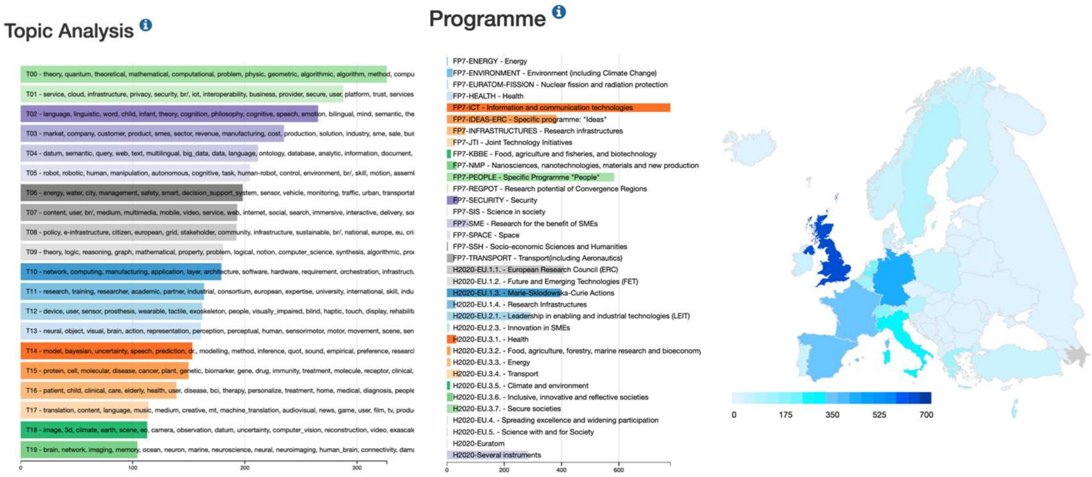
The development of Corpus Viewer started in 2016 and is still progressing based on the collaboration of several subcontracted University research groups and companies. Corpus Viewer on its 1.0 version, is currently used by three public administrations: SEAD (Ministry of Economy), the Spanish Foundation for Science and Technology (FECYT) and the State Secretary for University and Research, Development and Innovation (SEUIDI) at the Spanish Ministry of Science.

Although Corpus Viewer is a generic platform that can be exploited with virtually any collection of text documents, the current deployment of the platform mainly hosts R&D related text corpora:

Corpus	Docs in corpus	Time Frame
Projects funded by the Spanish State Plan for R&D	110 K	2004 - 2016
European R&D projects (CORDIS)	78 K	1984 - 2018
American R&D projects (NSF)	150 K	1985 - 2017
American Health R&D projects (NIH)	1.8 M	1983 - 2017
Patent applications (PATSTAT)	90 M	1898 - 2017
Scientific papers with contributions from authors affiliated to a Spanish institution (SCOPUS)	680 K	2006 - 2018

These data sources are processed to assist in the definition and implementation of R&D&i public policies through a set of functionalities allowing to:

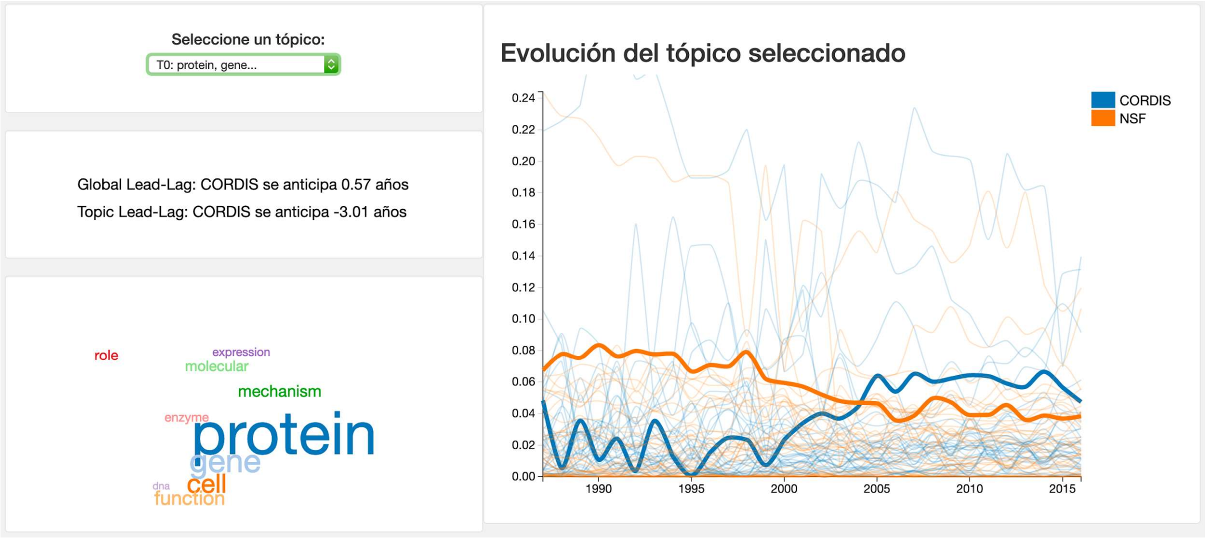
1. compare R&D&i funding and knowledge areas in different geographic regions,
2. identify competitive advantages between countries, regions, organizations,
3. identify R&D&i knowledge areas, as well as their emergence, evolution and even hybridization with other knowledge areas (it provides also metadata aggregation and BI type dashboard visualization),
4. R&D agent (organization, researcher and firm) profiling and,
5. assist in the assessment of the impact of public policies by tracking the outputs of grants, short and long term outcomes in terms of lead-lag.


Topic Analysis
T00 - theory, quantum, theoretical, mathematical, computational, problem, physic, geometric, algorithmic, algorithm, method, compu
T01 - service, cloud, infrastructure, privacy, security, br/, iot, interoperability, business, provider, secure, user, platform, trust, services
T02 - language, linguistic, word, child, infant, theory, cognition, philosophy, cognitive, speech, emotion, bilingual, mind, semantic, the
T03 - market, company, customer, product, smes, sector, revenue, manufacturing, cost, production, solution, industry, sme, sale, bus
T04 - datum, semantic, query, web, text, multilingual, big_data, data, language, ontology, database, analytic, information, document,
T05 - robot, robotic, human, manipulation, autonomous, cognitive, task, human-robot, control, environment, br/, skill, motion, asseml
T06 - energy, water, city, management, safety, smart, decision_support_system, sensor, vehicle, monitoring, traffic, urban, transportat
T07 - content, user, br/, medium, multimedia, mobile, video, service, web, internet, social, search, immersive, interactive, delivery, so
T08 - policy, e-infrastructure, citizen, european, grid, stakeholder, community, infrastructure, sustainable, br/, national, europe, eu, cri
T09 - theory, logic, reasoning, graph, mathematical, property, problem, logical, notion, computer_science, synthesis, algorithmic, pro
T10 - network, computing, manufacturing, application, layer, architecture, software, hardware, requirement, orchestration, infrastructu
T11 - research, training, researcher, academic, partner, industrial, consortium, european, expertise, university, international, skill, indu
T12 - device, user, sensor, prosthesis, wearable, tactile, exoskeleton, people, visually_impaired, blind, haptic, touch, display, rehabilit
T13 - neural, object, visual, brain, action, representation, perception, perceptual, human, sensorimotor, motor, movement, scene, sen
T14 - model, bayesian, uncertainty, speech, prediction, dr., modelling, method, inference, quot, sound, empirical, preference, researc
T15 - protein, cell, molecular, disease, cancer, plant, genetic, biomarker, gene, drug, immunity, treatment, molecule, receptor, clinical,
T16 - patient, child, clinical, care, elderly, health, user, disease, bci, therapy, personalize, treatment, home, medical, diagnosis, people
T17 - translation, content, language, music, medium, creative, mt, machine_translation, audiovisual, news, game, user, film, tv, produ
T18 - image, 3d, climate, earth, scene, eo, camera, observation, datum, uncertainty, computer_vision, reconstruction, video, exascal
T19 - brain, network, imaging, memory, ocean, neuron, marine, neuroscience, neural, neuroimaging, human_brain, connectivity, dam
0
100
200
300
Programme
FP7-ENERGY - Energy
FP7-ENVIRONMENT - Environment (including Climate Change)
FP7-EURATOM-FISSION - Nuclear fission and radiation protection
FP7-HEALTH - Health
FP7-ICT - Information and communication technologies
FP7-IDEAS-ERC - Specific programme: "Ideas"
FP7-INFRASTRUCTURES - Research infrastructures
FP7-JTI - Joint Technology Initiatives
FP7-KBBE - Food, agriculture and fisheries, and biotechnology
FP7-NMP - Nanosciences, nanotechnologies, materials and new production
FP7-PEOPLE - Specific Programme "People"
FP7-REGPOT - Research potential of Convergence Regions
FP7-SECURITY - Security
FP7-SIS - Science in society
FP7-SME - Research for the benefit of SMEs
FP7-SPACE - Space
FP7-SSH - Socio-economic Sciences and Humanities
FP7-TRANSPORT - Transport(including Aeronautics)
H2020-EU.1.1. - European Research Council (ERC)
H2020-EU.1.2. - Future and Emerging Technologies (FET)
H2020-EU.1.3. - Marie-Sklodowska-Curie Actions
H2020-EU.1.4. - Research Infrastructures
H2020-EU.2.1. - Leadership in enabling and industrial technologies (LEIT)
H2020-EU.2.3. - Innovation in SMEs
H2020-EU.3.1. - Health
H2020-EU.3.2. - Food, agriculture, forestry, marine research and bioeconomy
H2020-EU.3.3. - Energy
H2020-EU.3.4. - Transport
H2020-EU.3.5. - Climate and environment
H2020-EU.3.6. - Inclusive, innovative and reflective societies
H2020-EU.3.7. - Secure societies
H2020-EU.4. - Spreading excellence and widening participation
H2020-EU.5. - Science with and for Society
H2020-Euratom
H2020-Several instruments
0
200
400
600
0
175
350
525
700

Corpus Viewer also provides tools for policy implementation, in particular for the selection of evaluators or the retrieval of relevant documents (patents, scientific publications, R&D aid grants and proposals, etc) for innovation evaluation. Furthermore, it is used for plagiarism detection, identification of cases of double funding and fraud in aid grants and proposals submitted for national funding.

Corpus Viewer functionalities

1. Scalable NLP pipeline and Machine Translation of large volumes of documents.
2. Automatic classification of documents according to available taxonomies using deep learning networks.
3. Topic modeling analysis of document collections and topic inference for new documents.
4. Information retrieval system based document similarity.
5. Optimized columnar and textual indexing for efficient searches and query metadata filtering.
6. Tracking of semantically alike documents (semantic clusters) (emergence, evolution and hybridization with other clusters.
7. Dynamic topic analysis and temporal thematic evolution. Temporal analysis by areas of knowledge, lead-lag between different types of document corpus.
8. Topic-enhanced dashboards to analyze R&D&i distribution by metadata, including geographic area.
9. Automatic profiling and disambiguation of R&D&i key players based on their R&D production.
10. Analysis of collaboration networks



Seleccione un tópico:

T0: protein, gene...

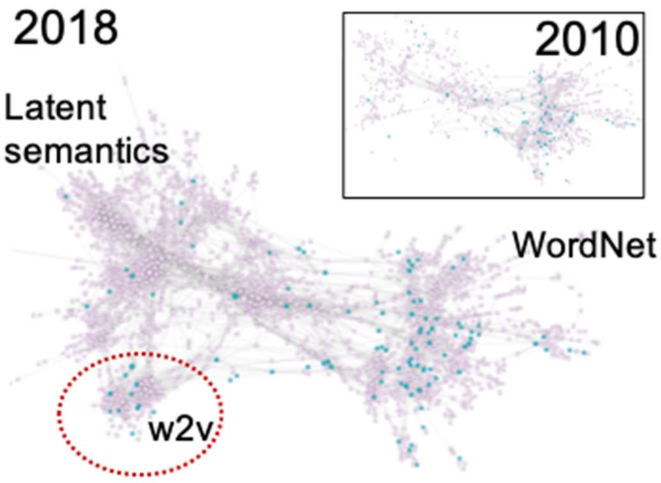
Global Lead-Lag: CORDIS se anticipa 0.57 años

Topic Lead-Lag: CORDIS se anticipa -3.01 años

Corpus Viewer Facilitating Techniques

Unlike traditional statistical description methods, our approach does not require predefined taxonomies or controlled vocabularies. Analysis techniques are based on underlying topics, conceptual indexing of documents, word Embeddings and other techniques of documentary representation.

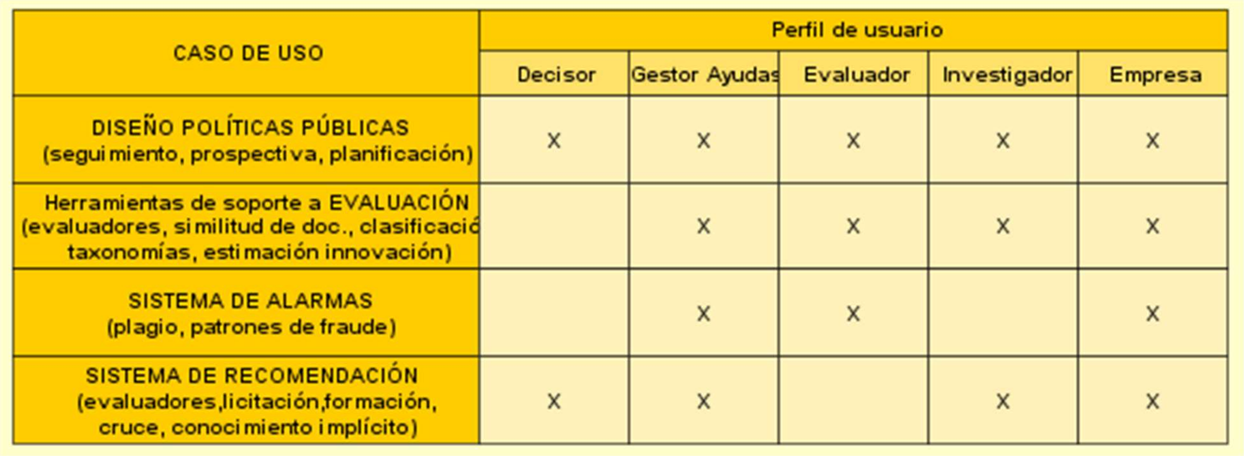
1. Scalable tokenization, PoS tagging, lemmatization, disambiguation and wikification for English and Spanish languages.
2. Automatic translation (ES-EN), NMT based.
3. Topic modeling, including: static models (Latent Dirichlet Allocation, LDA, CTM), dynamic topic models, hierarchical LDA, and recursive LDA
4. Textual search and document similarity implemented for topic, bag of words, and word Embeddings document representations
5. Analysis of graphs; modularity, distances between clusters and centrality calculation



Corpus Viewer Use Cases

Potential users of current Corpus Viewer platform include:

1. R&D policy makers
2. Managers and coordinators of R&D programs (policy implementation)
3. Grant evaluators
4. R&D agents: researchers, Public Research Institutions, Companies



CASO DE USO	Perfil de usuario				
	Decisor	Gestor Ayudas	Evaluador	Investigador	Empresa
DISEÑO POLÍTICAS PÚBLICAS (seguimiento, prospectiva, planificación)	X	X	X	X	X
Herramientas de soporte a EVALUACIÓN (evaluadores, similitud de doc., clasificació taxonomías, estimación innovación)		X	X	X	X
SISTEMA DE ALARMAS (plagio, patrones de fraude)		X	X		X
SISTEMA DE RECOMENDACIÓN (evaluadores,licitación,formación, cruce, conocimiento implícito)	X	X		X	X

Corpus Viewer Demonstrators

Corpus Viewer access is restricted to authorized users. However, within the project several demonstrators [Enlace a la Página web de demostradores de IntelComp] have been developed that provide access to part of the functionality implemented in the platform.