Boletín ANABAD. LXXII (2022), NÚM. 2, ABRIL-JUNIO. MADRID. ISSN: 2794-0519 (USB) - 2444-7293 (Internet)

INTELE : promoviendo la participación en las infraestructuras ERIC CLARIN y DARIAH

Autores:

Mikel Iruskieta, Centro HiTZ - Grupo Ixa (UPV/EHU). *Profesor Contratado Doctor en la Facultad de Educación de Bilbao. Departamento de Didáctica de la Lengua y la Literatura.*

Ainara Estarrona, Centro HiTZ - Grupo Ixa (UPV/EHU). *Investigadora postdoctoral contratada en la Facultad de Informática. Departamento de Lenguajes y Sistemas Informáticos.*

Aritz Farwell, Centro HiTZ - Grupo Ixa (UPV/EHU). *Investigador postdoctoral contratado en la Facultad de Informática. Departamento de Lenguajes y Sistemas Informáticos.*

German Rigau, Centro HiTZ - Grupo Ixa (UPV/EHU). *Profesor Titular de Universidad en la Facultad de Informática. Departamento de Lenguajes y Sistemas Informáticos.*

Dirección:

Universidad del País Vasco (UPV/EHU)

Mikel Iruskieta: mikel.iruskieta@ehu.eus

Fecha: 02/11/2022

Palabras clave: Infraestructura, Humanidades y Ciencias Sociales, CLARIN, DARIAH.

Key words: Infrastructure, Social Sciences and Humanities, CLARIN, DARIAH.

Resumen

La red de investigación estratégica INTELE ha sido financiada por el Ministerio de Ciencia, Innovación y Universidades de España (RED2018-102797-E), en el marco de las acciones de dinamización «Redes de Investigación» correspondientes al Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema de I+D+i, 2017-2020. Los objetivos principales de la red INTELE son impulsar actividades de promoción de las tecnologías del lenguaje para el castellano y las lenguas cooficiales (euskera, catalán y gallego) en CLARIN-ERIC y DARIAH-ERIC y lograr la participación oficial de España en las mismas. Sin la participación plena en dichas infraestructuras el conocimiento cultural y lingüístico de todas estas comunidades no va a estar representado adecuadamente y tampoco se podrán desarrollar los recursos y herramientas digitales en las infraestructuras europeas. En este trabajo explicaremos las actividades realizadas durante los dos años que ha durado el proyecto empezando por la promoción de las infraestructuras y la creación de una comunidad interesada en la participación y desarrollo de infraestructuras propias, hasta llegar a conseguir el compromiso de las instituciones para ser miembros de pleno derecho de las infraestructuras CLARIN-ERIC y DARIAH-ERIC. Finalmente, expondremos la propuesta final de la red INTELE que ha sido la creación de CLARIAH-ES y sus nodos, lo que significa la participación conjunta en ambas infraestructuras de forma coordinada.

Abstract

The INTELE strategic research network was supported by Spain's Ministry of Science, Innovation and Universities (RED2018-102797-E) within the framework of the revitalization actions, "Research Networks," corresponding to the State Program for the Generation of Knowledge and Scientific and

Technological Strengthening of the System of R+D+i, 2017-2020. The main objectives of the INTELE network are to achieve Spain's official participation in CLARIN-ERIC and DARIAH-ERIC, as well as to promote activities that further language technologies in these for Spanish and the co-official languages (Basque, Catalan, and Galician). Without full participation in these infrastructures, neither the cultural and linguistic knowledge of these communities will be adequately represented, nor will it be possible to develop digital resources and tools in the European infrastructures. In this paper, we will detail the activities carried out during the two years that the project lasted, beginning with the promotion of infrastructures and the creation of a community interested in the participation and development of its own infrastructures, followed by the eventual commitment of various institutions to become full members of the infrastructures CLARIN-ERIC and DARIAH-ERIC. To conclude, we will present the final result of the INTELE network : the creation of CLARIAH-ES and its nodes, which will participate jointly in both infrastructures in a coordinated manner.

INTRODUCCIÓN

España, hasta la fecha, no ha participado en las infraestructuras de investigación europeas CLARIN-ERIC (Váradi *et al.*, 2008) y DARIAH-ERIC (Kalman y Wandl-Vogt, 2014) relacionadas con las Humanidades y las Ciencias Sociales. Estas dos infraestructuras se encuentran ya plenamente operativas en muchos países de Europa, avaladas por la ESFRI y consolidadas como *Landmark* desde 2016. La ausencia de dichas infraestructuras de investigación tiene como consecuencia la falta de datos lingüísticos en castellano, euskera, gallego y catalán y también de recursos adecuados para investigar en dichas lenguas en las disciplinas de Humanidades y Ciencias Sociales. Desde la red estratégica INTELE (INfraestructura de TEcnologías del LEnguaje) hemos podido constatar que, para poder llevar a cabo su labor, los investigadores precisan de datos lingüísticos (entre otros) que puedan obtenerse fácilmente aún en una infraestructura dispersada físicamente pero centralizada digitalmente en una única página web, gracias al acceso confederado. La imposibilidad de recurrir a estas infraestructuras está dando lugar a que los investigadores pierdan peso en las publicaciones de estas disciplinas y no estén en las mismas condiciones para poder participar en proyectos competitivos

e innovadores, además de no poder realizar tareas de construcción de las infraestructuras.

INTELE es una red de investigación estratégica constituida inicialmente por investigadores españoles que están relacionados, por su participación anterior y su interés actual, con las infraestructuras europeas de investigación, ya constituidas como *European Research Infrastructure Consortium* o ERIC, para las Humanidades y Ciencias Sociales : CLARIN-ERIC (www.clarin.eu) y DARIAH-ERIC (www.dariah.eu) (en adelante CLARIN y DARIAH). El objetivo general de la red INTELE es impulsar actividades de promoción de las tecnologías del lenguaje en castellano y lenguas cooficiales (euskera, gallego y catalán), ya que en las infraestructuras europeas actuales no reside el conocimiento cultural y técnico de estas lenguas y por tanto su pretensión de reunir todos los recursos y herramientas de las lenguas y culturas europeas sería al menos parcial. Nuestra propuesta pretende llenar ese vacío en las infraestructuras europeas CLARIN y DARIAH. En este sentido, la red INTELE ha trabajado para lograr la participación oficial de España en dichas infraestructuras y seguir desarrollando ambas infraestructuras, a nivel europeo, del estado y de las comunidades autónomas. Pensamos que la participación de España como miembro de pleno derecho en estas infraestructuras contribuirá al avance de la investigación en Humanidades y Ciencias Sociales, así como a su posicionamiento estratégico en proyectos y programas internacionales, fundamentalmente en el contexto del Espacio Europeo de Investigación.

En este artículo presentaremos los resultados obtenidos por la red de investigación estratégica INTELE, así como sus actividades durante los dos años que ha durado el proyecto. El artículo está estructurado como sigue : tras esta introducción, en el apartado 2 haremos una presentación de la red INTELE y de su colaboración en las infraestructuras europeas CLARIN y DARIAH. A continuación, en el apartado 3 explicaremos en detalle las actividades llevadas a cabo por la red INTELE durante estos dos últimos años, y finalmente, en el apartado 4, miraremos al futuro presentando la propuesta de construcción de una infraestructura conjunta denominada CLARIAH-ES, que subsume la participación como miembros de pleno derecho en CLARIN y DARIAH y la coordinación de todos los nodos que crearán una infraestructura propia.

1. CLARIN, DARIAH, INTELE Y LA COLABORACIÓN EN AMBAS INFRAESTRUCTURAS

En este apartado resumimos las acciones que desde nuestro punto de vista han sido más relevantes en cuanto a la colaboración con las infraestructuras europeas CLARIN y DARIAH y cómo han confluido en la red estratégica INTELE, desde donde finalmente se ha logrado la participación oficial como miembro de pleno derecho en ambas infraestructuras con una propuesta conjunta : CLARIAH-ES.

1.1. El contexto europeo de las infraestructuras de investigación

Para explicar el contexto europeo de las infraestructuras de investigación que nos atañe en este artículo, es interesante recordar lo que Kaltenbrunner (2017) menciona sobre las iniciativas de creación de infraestructuras de investigación europeas. Las infraestructuras nacen desde la visión de la Comisión Europea sobre cómo desarrollar infraestructuras digitales; ya que en muchas ocasiones se han observado necesidades básicas similares (apoyo técnico, alojamiento de datos, servicios de computación o de análisis de datos, la interoperabilidad, la creación de comunidades y proyectos coordinados europeos, entre otros). Para el desarrollo de esa visión cabe recordar que todas las infraestructuras están organizadas por el mismo comité: ESFRI (*European Strategy Forum on Research Infrastructures*). Cuando estas iniciativas coordinadas digitalmente se consolidan, superan notablemente la fragmentación entre investigadores, entre proyectos de investigación y entre grupos de investigación.[1] Según Kaltenbrunner (2017), a diferencia de la organización de EEUU, la Comisión Europea considera que, sin esa coordinación, las instituciones académicas y los investigadores tienden a crear organizaciones que limitan el potencial del libre flujo de conocimiento. Desde este punto de vista, en la investigación sucede con demasiada frecuencia que para investigar se necesita rehacer muchas herramientas que ya están hechas, pero no son accesibles. Por lo tanto, la infraestructura digital se entiende como un medio para canalizar el diálogo entre las disciplinas más allá de las fronteras académicas, lingüísticas y geográficas. En las infraestructuras se promueve la reutilización de recursos y herramientas de investigación digital

1 https ://www.esfri.eu/esfri-white-paper

accesibles en abierto, para impulsar que los investigadores aborden métodos digitales compartidos e interoperables que permiten cambiar las preguntas de investigación y se consolidan temas de estudio innovadores.

Kaltenbrunner (2017) menciona que las raíces de la estrategia europea de infraestructura digital se remontan a los intentos de crear infraestructuras para las Humanidades a nivel estatal. En el Reino Unido, por ejemplo, esos primeros esfuerzos buscaban preservar, catalogar y fomentar el uso de los recursos digitales procedentes de la investigación en Humanidades. Un proyecto inicialmente llevado a cabo por el Servicio de Datos sobre Artes y Humanidades (AHDS), administrado por el King's College, fue financiado en parte por el Consejo de Investigación de las Artes y las Humanidades (AHRC). El AHRC dejó de financiar el proyecto doce años después, cuando decidió que las universidades británicas eran ya capaces de mantener servicios de datos digitales por sí mismas. En Alemania, otro ejemplo, la primera infraestructura digital se vio enriquecida por varios proyectos, y entre ellos, muy especialmente, por TextGrid en 2006 (Gietz *et al*., 2006). Cuando no existía una infraestructura estatal para las Humanidades Digitales, desde el proyecto TextGrid se animó a los investigadores de las universidades alemanas a que utilizaran su software y sus herramientas. Más tarde, en 2016, TextGrid empezó a participar en DARIAH.

Una cuestión que no se puede dejar a un lado y que ha sido cuestión de debate en los diferentes foros en los que la red estratégica INTELE ha participado ha sido la de la financiación. Es notable que estas propuestas de infraestructura necesitan de financiación y que dicha financiación no es soportable para los grupos de investigación. Aunque cada infraestructura se compone de forma diferente (algunas compuestas con agencias estatales, otras con la participación de grupos de investigación, investigadores o universidades), la financiación de CLARIN y DARIAH en los diferentes estados procede principalmente de fuentes estatales (y federales o autonómicas, en algunos casos). En este sentido, hay que tener en cuenta que la cada vez mayor dependencia de las inversiones y de la financiación basada en proyectos competitivos de investigación de corto plazo es un problema importante para las infraestructuras digitales que quieren desarrollarse, mantener datos y herramientas en el tiempo para ofrecer mejores servicios a los investigadores. Además, se debe considerar que si estas infraestructuras participan en dichos proyectos competitivos

y acaparan las fuentes de financiación, la comunidad investigadora puede sentir una competencia desleal y esta comunidad puede dejar de participar o aportar en la infraestructura. En este contexto, la Comisión Europea se ha convertido en una importante fuente de financiación y apoyo político, como se ha visto en programas como FP7, Horizon 2020 y los Fondos Estructurales europeos.

En cuanto a los objetivos del Espacio Europeo de Investigación (EEI), Kaltenbrunner (2017) menciona que la Comisión Europea confía en que estas organizaciones supra-estatales digitales y coordinadas que ofrecen acceso libre al conocimiento científico, ayudarán a los investigadores a cambiar de paradigma metodológico y así poder realizar investigaciones innovadoras sobre nuevos temas utilizando métodos digitales interoperables. De esta forma las infraestructuras y la Comisión Europea tratan de superar la fragmentación geográfica, epistémica e institucional (es decir, uno de los objetivos del Espacio Europeo de Investigación (EEI)). Es indudable que la creación del Espacio Europeo de Investigación (EEI) y de las infraestructuras digitales, CLARIN y DARIAH entre otras, han aumentado los recursos, mejorado las herramientas e incrementado el impacto y los resultados científicos de la investigación en la sociedad.

El anteriormente mencionado Foro Estratégico Europeo sobre Infraestructuras de Investigación (ESFRI) desempeña un papel fundamental en la gestión de los planes de infraestructuras digitales, pero cabe señalar que no tiene la capacidad de sustituir la política nacional de los países europeos. La ESFRI tiene por objeto la coordinación de las inversiones en infraestructuras digitales transeuropeas y la dirección de su hoja de ruta, diseñada para identificar nuevas y significativas infraestructuras a escala europea. Para su inclusión en la hoja de ruta una infraestructura debe haber llegado a un nivel de madurez adecuado, recibiendo una mención (*Landmark*), como se ha señalado anteriormente. La hoja de ruta de la ESFRI marca el principio y el final de un proyecto de infraestructura. La Comisión Europea financia los proyectos de infraestructuras en una fase inicial, pero los Estados miembros son responsables de la mayoría de los costes durante la fase de ejecución. Esto se lleva a cabo a través del Consorcio de Infraestructuras de Investigación Europeas (ERIC por su nombre original en inglés : *European Research Infrastructure Consortium*), que dota a los proyectos de personalidad

jurídica, lo que les permite participar en convocatorias europeas y nacionales. Las subvenciones van a proyectos preexistentes en varios países, pero únicamente cuando dichos proyectos se ajustan al enfoque de la política de la Comisión Europea. La coordinación transnacional es obligatoria, por lo que las infraestructuras deben organizar el desarrollo de recursos y herramientas de acuerdo con las prácticas de investigación financiadas.

CLARIN y DARIAH, infraestructuras con el Landmark de la ESFRI, aspiran a erigir recursos y herramientas de investigación completas para una variedad lo más amplia posible de disciplinas y lenguas, pero sobre todo para las Humanidades, Artes y Ciencia Sociales.

1.2. CLARIN-ERIC

El objetivo de CLARIN (Krauwer y Hinrichs, 2014) es organizar centros de investigación europeos para hacer accesibles todos los recursos y herramientas digitales del lenguaje. La comunidad investigadora a la que se quiere dar acceso a estos recursos y herramientas es la de las Humanidades y las Ciencias Sociales. Desde el 2012 CLARIN está proporcionando un acceso fácil y sostenible a datos lingüísticos digitales (texto escrito, voz o multimodal) de la infraestructura. El uso de la infraestructuras se puede ver en las publicaciones de CLARIN donde se explica la experiencias de los usuarios[2].

Es importante mencionar que la motivación de proporcionar estos datos y herramientas no es un fin en sí mismo, la motivación es empoderar a la comunidad científica para que realice su investigación con las herramientas del siglo XXI, investigación abierta y con mayor repercusión en la sociedad. Dicho de otra forma, la motivación es ayudar en el cambio de paradigma metodológico de las Humanidades y Ciencias Sociales (Iruskieta y Bel, 2017), utilizando herramientas avanzadas que permitan : i) considerar una mayor cantidad de datos y poder responder a nuevas preguntas de investigación, ii) mejorar la recogida de datos y reducir el tiempo en esa tarea, iii) ayudar en el mantenimiento de datos, recursos, visibilidad de la investigación, iv) ofrecer consultoría técnica, facilita la planificación y v) crear o impulsar las buenas prácticas. Por ejemplo, CLARIN ofrece herramientas avanzadas

2 https ://zenodo.org/record/4288980#.X9YDS7N7mDI.

para explorar, explotar, anotar, analizar o combinar dichos conjuntos de datos lingüísticos en la infraestructura, aunque los datos estén en centros o universidades diferentes. Esta coordinación de centros de investigación se organiza gracias a la federación de centros CLARIN organizados en red y al acceso confederado de los investigadores en cualquiera de los recursos que ofrecen estos centros de investigación. La infraestructura CLARIN está funcionando plenamente en la mayoría de países europeos y gracias a que una gran cantidad de centros participantes están ofreciendo todos estos servicios.

Pero la cobertura de la infraestructura es limitada si no participan algunos países, centros o comunidades científicas. Un ejemplo de ello es la cobertura por lenguas de los recursos CLARIN. En la Virtual Language Observatory[3] donde se pueden explorar los datos y herramientas que hay en la infraestructura, que hay más de 1 millón de componentes para lenguas diferentes[4] : inglés (154.928), alemán (143.271), holandes (117.312), danes (109.962), esloveno (73.495), polaco (40.466), francés (24.432) y africano (7.870). Aun sin la participación oficial de España, los recursos en las lenguas del estado son los siguientes : castellano (14.444); catalan, valenciano (1.364), euskera (498), gallego (216). Esos recursos son de modalidades diferentes : recursos para el habla (99.558), texto escrito (1.543), gestos (1.307), expresiones faciales (452), estados emocionales (451).

En cuanto a los servicios que ofrece CLARIN, observamos que algunas herramientas se desarrollan a modo general, pero que hay otras muchas herramientas que se podrían desarrollar en castellano y en lenguas cooficiales y no están desarrolladas en la infraestructura (ver Tabla 1) :

	PL	**DE**	**EN**	**ES.**
1. Constituency Parsing		x	x	
2. Coreference Resolution	x			
3. Dependency Parsing	x	x	x	x
4. Distant Reading	x	x	x	x
5. Extraction of Polish terminology	x			

3 https://vlo.clarin.eu

4 La cantidad de los datos puede evolucionar en la infraestructura, ofrecemos los datos para ejemplificar los recursos y el tamaño de la infraestructura.

6. Inclusion detection	x			
7. Keyword Extractor	x			
8. Lemmatization		x	x	
9. Machine Translation		x	x	
11. Morpho-syntactic tagger	x		x	
12. Morphological Analysis	x	x	x	
13. Named Entity Recognition	x	x	x	x
14. Named Entity Relation Detection				
15. Part-Of-Speech Tagging	x	x	x	
16. Sentiment Analysis	x			
18. Spatial expression detection	x			
19. Speech Recognition				
20. Stylometry				
28. Word sense disambiguation	x			

Tabla 1 : *Servicios interoperables para el análisis de texto escrito en diferentes lenguas marcadas con una "x" que ofrece CLARIN con la Switchboard de CLARIN*

Es de mencionar, además, todos los servicios que ofrece CLARIN (*BAS Web Services*)[5] para muchas lenguas para el análisis de texto hablado : Constituency Parsing, Coreference Resolution, Dependency Parsing, Distant Reading, Extraction of Polish terminology, Inclusion detection, Keyword Extractor, Lemmatization, Machine Translation, Metadata Processing, Morpho-syntactic tagger, Morphological Analysis, Named Entity Recognition, Named Entity Relation Detection, Part-Of-Speech Tagging, Sentiment Analysis, Shallow Parsing, Spatial expression detection, Speech Recognition, Stylometry, TF, IDF, TF-IDF calculation, Text Analytics, Text Enhancement, Text Summarization, Tokenization, Topic Modelling, Visualisation of Geographic Data y Word sense disambiguation.

1.3. DARIAH-ERIC

En cambio el objetivo de DARIAH (Romary, 2014) es el de investigar y enseñar este cambio de paradigma digital en las disciplinas del área de las Humanidades y el Arte, como los estudios literarios, la historia y la filosofía. Por tanto, la infraestructura impulsa los grupos de trabajo, relacionando investigadores de diferentes estados y de diferentes disciplinas. Estas

5 https ://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

disciplinas tienen enfoques metodológicos y teóricos diversos, y cuentan a su vez con diferentes tradiciones nacionales de investigación. Esto, junto a la falta de un compromiso firme con las herramientas digitales para las Humanidades, sugiere que el deseo de la Comisión Europea de unificar la investigación de diversas disciplinas a lo largo y ancho del paisaje europeo será difícil de alcanzar aun mediante las infraestructuras digitales. Para lograr ese objetivo, previamente las Humanidades deben desarrollar una relación más estrecha con las tecnologías y los recursos digitales. Además, hay quien duda de la idoneidad de un enfoque centralizador en las infraestructuras digitales (por ejemplo, al que se ha seguido en proyectos como el CERN), sea el más adecuado para las Humanidades. De hecho, en DARIAH se rehúye de este planteamiento centralizador y los responsables de las infraestructuras deben ser gestores a tiempo completo, ya que consideran que los investigadores de las infraestructuras deben realizar sus investigaciones e interactuar con sus comunidades científicas. Por tanto, en este caso su visión de las infraestructuras es más descentralizada, y se subraya la coordinación y la integración de los instrumentos que ya han sido creados por instituciones de toda Europa. Así, la gestión de las herramientas en la infraestructura está en manos de los investigadores creadores (si es que así lo desean). La estrategia de DARIAH es buscar una mayor integración, mientras al mismo tiempo se reconoce la diversidad de temas y métodos de investigación. La infraestructura digital se organiza en grupos metodológicos en los que se incorporan herramientas que pueden ser utilizadas por diferentes disciplinas. Estos son los grupos de trabajo activos de DARIAH : Combining Language Learning with Crowdsourcing Techniques (D4COLLECT), Ethics and Legality in the Digital Arts and Humanities (ELDAH), Digital Practices for the Study of Urban Heritage (UDigiSH), Bibliographical Data (BiblioData), Theatralia, Research Data Management, Artificial Intelligence and Music (AIM), #dariahTeach, Digital Numismatics, DH Course Registry, Guidelines and Standards, Visual Media and Interactivity, Analyzing and linking biographical data, Thesaurus Maintenance, GeoHumanities, Mediaevalist's Sources (MESO), Sustainable publishing of (meta)data, Lexical Resources, Women Writers in History y Image Science and Media Art Research.

1.4. Principios FAIR y CARE para desarrollar las infraestructuras

Todas estas iniciativas con sus recursos y herramientas tienen como objetivo no solo el de aumentar el impacto de la investigación sino de hacer ciencia abierta siguiendo unos principios, como pueden ser los principios FAIR y CARE.

Por ejemplo, se puede decir que la infraestructura CLARIN se desarrolló en los principios FAIR, *avant la lettre*, antes de que fueran propuestos. En 2016, los "Principios rectores de FAIR para la gestión y administración de datos científicos" se publicaron en *Scientific Data*. La intención de esa publicación era la de proporcionar pautas para mejorar la Encontrabilidad, Accesibilidad, Interoperabilidad y Reutilización de los activos digitales. Dichos principios enfatizan la capacidad de acción de las máquinas (es decir, la capacidad de los sistemas computacionales para encontrar, acceder, interoperar y reutilizar datos con una intervención humana mínima y sencilla) porque los humanos confían cada vez más en el soporte computacional para manejar los datos como resultado del aumento en el volumen, complejidad y velocidad de creación o de uso de datos.[6]

Hacer ciencia de forma abierta y justa incluye el derecho a crear valor a partir de los datos propios de manera que se base en las cosmovisiones propias o autóctonas y aprovechar las oportunidades dentro de la economía del conocimiento. Los Principios de CARE para la Gobernanza de Datos Autóctonos están orientados a personas y propósitos, lo que refleja el papel crucial de los datos en el avance de la innovación y el empoderamiento de los datos propios creados en la comunidad. Estos principios CARE complementan los principios FAIR existentes que alientan el movimiento de datos abiertos y de otro tipo para considerar tanto a las personas como a los propósitos en la defensa de las comunidades y de sus actividades.[7]

1.5. Relación de las ERICs CLARIN y DARIAH y la red INTELE

Catalogar todas las iniciativas digitales de la investigación en Humanidades y Ciencias Sociales es una tarea ardua que difícilmente se puede realizar sin la participación de los grupos de investigación o los investigadores. Sin embargo, mencionar la participación en las infraestructuras europeas es una

6 https ://www.go-fair.org/fair-principles/
7 https ://www.gida-global.org/care

tarea más concreta. Cabe destacar que las infraestructuras de investigación tienen como objetivo conectar estas iniciativas digitales y dar servicios de todo tipo sobre tareas del ciclo de investigación. La red INTELE ha logrado desde el principio reunir a las iniciativas más significativas relacionadas con dichas infraestructuras :

- 2007-2013. Creación del centro de competencias CLARIN (IULA-UPF-CC-CLARIN) cofinanciado por el programa FEDER de Cataluña (2007-2013).[8]

- 2016. Creación del primer centro CLARIN *Knowledge* (CLARIN-K) europeo o centro de conocimiento lingüístico CLARIN en España (Bel *et al*., 2016) con 4 nodos CLARIN-K :[9]

    - IULA-UPF (Barcelona) CLARIN *Competence Center* IULA-UPF (CC-CLARIN IULA-UPF) especializado en analíticas de texto y tecnologías del lenguaje.

    - IXA-UPV/EHU (Donostia) especializado en analíticas de texto y tecnologías del lenguaje para el euskera y otras lenguas.[10]

    - LINDH-UNED (Madrid) Laboratorio de Innovación en Humanidades Digitales (LINDH) centro para el desarrollo de las Humanidades Digitales en la UNED.[11]

    - TALG Group (Vigo) especializado en el desarrollo de tecnologías del lenguaje para promover y facilitar el uso del gallego y ahora este nodo gallego lo dirige el Instituto de la Lingua Gallega (USC).[12]

- 2017-2020. Participación en la comisión *Knowledge Sharing Infrastructure* (KSI) de CLARIN, que se encarga de la formación, difusión del conocimiento y de las experiencias de los centros CLARIN-K.

8 http ://www.clarin-es-lab.org/index-es.html
9 http ://ixa2.si.ehu.eus/clarin-es/
10 http ://ixa2.si.ehu.eus/clarink/index.php?lang=es
11 https ://linhd.uned.es/clarin-centre-k/
12 https ://ilg.usc.gal/clarin-center/

- 2019. Creación de IMPACT,[13] el segundo centro de conocimiento CLARIN en digitalización (Universidad de Alicante).[14]

- 2020-2022. Red estratégica INTELE financiada por el Ministerio Ciencia, Innovación y Universidades de España (RED2018-102797-E) coordinado por la Universidad del País Vasco (UPV/EHU), en el marco de las acciones de dinamización «Redes de Investigación» correspondientes al Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema de I+D+i, 2017-2020.[15]

- 2022, agosto. Participación en DARIAH, como Cooperating Partner de LINDH-UNED.[16]

- 2022, septiembre. Propuesta de infraestructura CLARIAH-ES (participación oficial de España en las infraestructuras europeas CLARIN-ERIC y DARIAH-ERIC) con al menos 10 nodos confirmados.

Una vez explicado el contexto en el que trabajamos, en la siguiente sección detallamos las actividades más importantes de la red estratégica INTELE para poder llegar a la propuesta conjunta de CLARIAH-ES.

2. ACTIVIDADES DE LA RED ESTRATÉGICA INTELE

La red estratégica INTELE ha desarrollado diversas actividades durante los dos años en los que ha estado vigente. En las siguientes secciones detallaremos dichas actividades empezando por los workshops, siguiendo por la presentación del manifiesto INTELE y terminando por los webinarios.

2.1. I Workshop INTELE de presentación

Con el principal objetivo de impulsar actividades de promoción de las

13 https ://www.digitisation.eu/helpdesk/
14 https ://www.clarin.eu/news/impact-clarin-k-centre-digitisation-impact-ckc
15 http ://ixa2.si.ehu.eus/intele/
16 https ://www.dariah.eu/2022/08/04/linhd-joins-dariah-as-cooperating-partner/

infraestructuras europeas de investigación CLARIN y DARIAH para lograr la participación oficial de España en las mismas, la red INTELE, dada la situación de pandemia mundial, celebró en formato virtual su primer workshop de presentación el día 20 de noviembre de 2020.

Para formar parte como miembros con pleno derecho de ambas infraestructuras, el Gobierno de España necesitaba tener evidencias del interés de los investigadores en utilizar, colaborar, participar y compartir herramientas digitales, recursos electrónicos y proyectos relacionados con las tecnologías del lenguaje y la digitalización.

Por ese motivo, era necesario un esfuerzo cooperativo para recabar información individual o por grupos de investigadores, y de estudios de textos en cualquiera de las lenguas del Estado y sus variantes. Se invitó a quienes quisieran formar parte de este proyecto de interés común para la comunidad científica a participar en este primer workshop para tener una aproximación de la masa crítica con la que se podía contar.

Este primer workshop contó con mesas redondas en las que participaron tanto representantes políticos del Gobierno Vasco como destacados especialistas de diferentes ámbitos y se trataron los siguientes temas :

- Políticas de apoyo a infraestructuras digitales europeas para las Humanidades y las Ciencias Sociales.
- Las Humanidades Digitales en España : potencial y necesidad.
- Participación y difusión de proyectos y recursos de grupos de investigación de Humanidades y Ciencias Sociales.

Tanto el programa del workshop, como las presentaciones de los diferentes ponentes se pueden consultar en la página web de INTELE : https://ixa2.si.ehu.eus/intele/actividades

La participación en este primer workshop fue todo éxito con más de 150 inscripciones lo que daba clara muestra del interés que existe en la comunidad investigadora sobre las infraestructuras europeas CLARIN y DARIAH.

Cabe señalar también que se invitó a los participantes a enviar pósteres de presentación de los diferentes proyectos o grupos de investigación del área de las Humanidades y Ciencias Sociales. Se recibieron 42 pósteres en total y se pueden consultar en este enlace : http://ixa2.si.ehu.eus/intele/posters

Después de este primer workshop pudimos constatar el interés de muchos grupos de investigación e investigadores individuales en participar en la red INTELE por lo que abrimos un proceso de adhesión a la red donde los diferentes grupos, instituciones o investigadores podían unirse como miembros participantes a la red. Tras este proceso de adhesión la red cuenta hoy en día con 156 grupos de investigación como miembros participantes o colaboradores. Puede verse el listado de todos los grupos participantes en este enlace : http://ixa2.si.ehu.eus/intele/miembros_participantes

Además, en los últimos 6 meses 3 nuevos organismos se han incorporado a las actividades de la red : el CSIC, el Barcelona Supercomputing Center (BSC-CNS) y el Instituto da Lingua Galega (USC).

2.2. II Workshop INTELE de internacionalización

Una vez constatado el interés de los investigadores en impulsar la iniciativa y participar en la red y después del éxito del primer workshop se decidió organizar un segundo workshop el 1 de julio del 2021, también en formato virtual, con el objetivo de definir nuestras necesidades para la promoción y participación en las infraestructuras internacionales y paneuropeas.

Al comienzo del workshop se ofreció información sobre los avances de la Red INTELE y sobre cómo participar en sus actividades. Además se dieron a conocer los resultados de los contactos mantenidos con las administraciones responsables de la financiación de las infraestructuras de investigación a nivel autonómico y estatal.

No obstante, la mayor parte de la reunión fue una sesión de trabajo durante la cual se trabajó en 5 áreas por grupos con el objetivo de discutir la situación y las necesidades de cada área. Los grupos de trabajo (GT) constituidos *ad hoc* fueron los siguientes : i) GT1 : Historia, Sociología, Comunicación,

Musicología; ii) GT2 : Lengua, Lingüística, Sociolingüística, Traducción, Literatura, Teatro, Filología; iii) GT3 : Biblioteconomía, Museología, Arte, Arqueología, Arquitectura; iv) GT4 : Educación, y v) GT5 : Derecho.

El objetivo principal de esta reunión era hacer una lista entre los participantes de las necesidades y los retos futuros de la investigación española en las disciplinas de Humanidades y Ciencias Sociales que trabajan con textos y en el marco de la necesaria digitalización de métodos y procesos. La misión era elaborar entre todos un ránking fundamentado de las infraestructuras digitales que serán el apoyo necesario para los investigadores en España, o al menos un apoyo igual a los investigadores europeos que les capacitan para entrar en consorcios y acceder a fondos europeos.

Así pues, en este segundo Workshop reunimos a 152 investigadores, expertos y responsables de la investigación de diferentes instituciones para obtener una visión informada de las necesidades ahora y a cinco años vista.

La metodología de trabajo estaba dirigida a descubrir que es una infraestructura y conocer cuáles son las necesidades actuales y futuras de los investigadores y para ello se planteó un ejercicio en cada grupo de trabajo que consistía en hacer un listado ponderado de los 5 recursos/herramientas/servicios que querrían tener para la investigación en su ámbito, para que al final se pudiese realizar un decálogo de necesidades a las que la futura infraestructura debería responder. Cada grupo trabajó de la mano de dos dinamizadores, un miembro de la red INTELE y otra persona referente en su área.

Tras las sesiones de trabajo identificando necesidades, observamos que muchas de las carencias que los investigadores señalaban se repetían en los diferentes grupos. Una cosa que muchos investigadores en los diferentes grupos señalaron como imprescindible fue que la infraestructura debería presentarse desde las **necesidades de la investigación** y no desde lo digital o informático, es decir, debe organizarse según las necesidades de la investigación.

A continuación presentamos un resumen con las demandas prioritarias y más transversales :

1. **Almacenamiento** de contenidos (fuentes audiovisuales, orales, musicales, textuales y digitales), protección de datos y conservación de la investigación. Servicios para la preservación y migración. Apoyo para actualización de software para bases de datos en formatos antiguos y para la migración de datos a formatos nuevos. Una plataforma europea para bases de datos comunes cuya caducidad no dependa de la financiación.

2. **Servicios de formación** adaptados a diferentes niveles y perfiles. Oferta formativa y de recursos; en idioma castellano y con conceptos asequibles para todos los niveles formativos.

3. **Marco regulatorio** claro y amplio, inteligible. Asesoría legal.

4. **Servicios técnico-científicos** (contratación de personal). Infraestructuras generales para todos, de manera que el tiempo que se gaste en la parte administrativa la puedan realizar técnicos contratados con ese fin y aligerar, de esta manera, la carga del investigador.

5. **Servicios de asesoramiento técnico-digital** adaptados a diferentes niveles y perfiles. **Centros de competencias** prestando (o diseminando) a bajo coste servicios (e.g. buenas prácticas, etc) para instituciones patrimoniales pequeñas que por falta de escala o de conocimientos no abordan proyectos digitales. **Nodos de investigación emergente :** (1) Protocolos de metodologías de investigación "alternativas" (2) Aglutinar la investigación-acción (de docentes) para congregar esfuerzos de investigación (que puedan ser explotados por docentes y por grupos de investigación) y proporcionar servicios para investigaciones a pequeña escala (profesorado haciendo pequeñas investigaciones en aula).

6. **Servicios de computación para procesar y analizar textos** literarios, lingüísticos, comunicación digital disponible en internet, etc. Herramientas para analizar materiales multimodales. Incluir herramientas que utilizan la inteligencia artificial y procesamiento del lenguaje natural.

7. **Introducir la analítica de datos**. Anonimizar datos y contextualizar.

8. **Adaptar PLN básico y avanzado a diferentes idiomas y dominios.**
9. **Acceso a datos,** a los textos jurídicos de forma masiva, abierta y en formatos accesibles para los investigadores. Un repositorio integrado e interoperable (y relacionados) de colecciones digitales en acceso abierto de instituciones españolas/europeas para uso en computación
10. **Catálogos de servicios y recursos**.

2.3. Manifiesto INTELE

Durante el segundo workshop de INTELE también se pidió a los investigadores que pensaran una lista de acciones que ellos o sus respectivas instituciones podrían llevar a cabo para apoyar la incorporación de España a CLARIN y DARIAH. Y en ese sentido la red INTELE presentó al final del workshop un manifiesto de adhesión a la iniciativa para recabar el apoyo tanto de investigadores a título personal, como de instituciones u organismos que pudieran estar interesados.

Tenemos que señalar que el manifiesto ha resultado ser un éxito ya que ha recabado el apoyo de más 750 investigadores a título personal y más de 170 instituciones, entre las que se encuentran 15 universidades, 16 facultades, 14 institutos, 7 centros o facultades, 4 grupos de investigación, 3 departamentos, 2 asociaciones y 3 empresas o fundaciones, la RAE y algunas comunidades autónomas. Se puede consultar el listado de instituciones firmantes en este enlace : https://ixa2.si.ehu.eus/intele/node/88/webform-results/table

A continuación transcribimos aquí el manifiesto[17] en su integridad :

MANIFIESTO INTELE

Los investigadores, grupos y centros de investigación, universidades e instituciones abajo firmantes manifiestan lo siguiente :

- en el actual contexto de transformación digital, el estudio, la

17 El manifiesto INTELE puede consultarse en la página web (https ://ixa2.si.ehu.eus/intele/manifiesto) donde también existe un formulario para adherirse al mismo.

investigación y el desarrollo en Humanidades, Artes y Ciencias Sociales requieren infraestructuras tecnológicas que permitan el tratamiento computacional de datos textuales, visuales, numéricos y/o sonoros;

- estas infraestructuras fomentan el multilingüismo, la interoperabilidad, el mantenimiento y la reutilización de recursos, la ciencia abierta, la visibilidad y la cooperación científica en Europa; permiten superar así la fragmentación de las comunidades de investigación aumentando el impacto de su investigación;
- en Europa existen infraestructuras distribuidas de investigación de este tipo como CLARIN (centrada en datos y procesos digitales relacionados con el Lenguaje) y DARIAH (centrada en datos y procesos digitales relacionados con las Humanidades Digitales);
- a diferencia del resto de países de la Unión Europea, España no es miembro oficial de ninguna de estas infraestructuras y, por tanto, la investigación española no está desarrollando todo su potencial en igualdad de condiciones, además de quedar excluida de sus servicios y de la participación con tecnología propia;
- nuestras lenguas, culturas y realidades son fundamentales en cualquier investigación que se realice actualmente en estas áreas a escala europea;
- la contribución y participación española en ambas infraestructuras de investigación impulsará el desarrollo de las Humanidades Digitales, así como el posicionamiento estratégico en proyectos y programas internacionales en el contexto del Espacio Europeo de Investigación.

Por todo ello, solicitamos a las instituciones competentes la incorporación, desarrollo e impulso en España de las infraestructuras distribuidas de investigación europeas CLARIN y DARIAH

2.4. III Workshop INTELE final

Como punto final del trabajo de la red los días 13 y 14 de septiembre de

2022 se organizó el tercer workshop en el Ayre Gran Hotel Colón de Madrid, financiado por CLARIN y el Ministerio Ciencia e Innovación de España. El objetivo principal del taller fue presentar los resultados de la red INTELE y la propuesta de participación de España en las infraestructuras CLARIN y DARIAH del programa ESFRI de la Unión Europea. En este contexto INTELE dio a conocer la propuesta de CLARIAH-ES (participación conjunta de España en CLARIN y DARIAH) a los investigadores de las áreas de las Humanidades y Ciencias Sociales en España.[18]

Dada la importancia estratégica del evento y el hito que representa formar parte de las infraestructuras de investigación europeas CLARIN y DARIAH, el taller contó con la presencia de destacados representantes de las administraciones públicas que apoyan la iniciativa, así como con la participación de expertos de los ámbitos científico y académico. La inauguración del taller estuvo presidida por Raquel Yotti (Secretaria General de Investigación del Ministerio de Ciencia e Innovación), y contó también con la presencia de Cristina Gallach (Comisionada Especial para la Alianza por la Nueva Economía de la Lengua), Ana Isabel Cremades (Directora General de Investigación e Innovación Tecnológica de la Comunidad Autónoma de Madrid), Juncal Gutiérrez (Vicerrectora de Euskera, Cultura e Internacionalización de la UPV/EHU) y Eneko Agirre (Director del centro HiTZ). Todos destacaron la necesidad y oportunidad de formar parte de ambas infraestructuras y la Secretaria General de Investigación del Ministerio de Ciencia e Innovación, **Raquel Yotti, anunció la voluntad del Ministerio de participar formalmente en ambas infraestructuras a partir del próximo año 2023**.

Se puede consultar el resto del programa del taller en este enlace : https://ixa2.si.ehu.eus/intele/workshop_final

A este último workshop se inscribieron más de 120 personas y se enviaron un total de 33 pósteres que se pueden consultar aquí : https://ixa2.si.ehu.eus/intele/posters_iii_workshop

La clausura del taller contó con las intervenciones de José Manuel Pingarrón (Secretario General de Universidades) y Guillermo López Gallego (Subdirector General del Español en el Mundo) que subrayaron la

18 Steurs (2017) explica las infraestrcutruas de CLARIN, CLARIAH y DARIAH.

importancia y alcance de participar como miembro de pleno derecho en las infraestructuras CLARIN y DARIAH a partir del próximo año 2023.

2.5. Webinarios

Además de los workshops, durante el 2021 la red organizó diversos webinarios. La participación de los investigadores ha puesto de manifiesto el interés de la comunidad en la digitalización de los métodos de investigación.

Hemos intentado que todos los webinarios tuvieran una parte práctica donde los investigadores pudieran aprovechar para conocer diferentes herramientas o servicios de primera mano. El primer webinario de 5 tuvo lugar el 23 de febrero del 2021 y el último el 11 de junio del mismo año.

Estos han sido los webinarios organizados :

- 23 de febrero del 2021 : Proyecto ParlaMint. Ponentes : Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences), Petya Osenova (Sofia University, IICT-BAS) y María Calzada Pérez (Departamento de Traducción y Comunicación, Universitat Jaume I, Castellón de la Plana).

- 25 de marzo del 2021 : Programming Historian : Un proyecto colaborativo para poner la programación al alcance de los humanistas". Ponentes : Jennifer Isasi (The Pennsylvania State University) y Riva Quiroga (Universidad Católica de Chile).

- 17 de mayo del 2021 : "Text Analysis for Spanish". Ponente : Quinn Dombrowski (Stanford).

- 7 de junio del 2021 : "Facilitando el acceso computacional a colecciones digitales". (Biblioteca Virtual Miguel de Cervantes). Ponentes : Gustavo Candela (UA), María Pilar Escobar (UA) y María Dolores Sáez (UA).

- 11 de junio del 2021 : "Distant Reading". Ponentes : Christof Schöch (University of Trier, Germany), Borja Navarro (UA) y Rosario Arias (Universidad de Malaga).

Se puede consultar más información sobre los webinarios, tanto las

grabaciones y los recursos que se han creado para que puedan ser reutilizados en este enlace : https://ixa2.si.ehu.eus/intele/webinars

2.6. Reuniones con la administración y entidades referenciales

Por último queremos mencionar que la red INTELE ha realizado numerosas reuniones con distintos agentes para presentar la propuesta CLARIAH-ES, entre otras: i) con la administración del estado, ii) con representante de varias comunidades autónomas, iii) con ministerios, iv) con representantes de la ESFRI, v) con los y las directoras de las infraestructuras europeas CLARIN y DARIAH, vi) con los vicerrectores de las universidades que han propuesto ser nodos de CLARIAH-ES, vii) con asociaciones de Humanidades Digitales, viii) con entidades de renombre y, finalmente, ix) con investigadores que han mostrado interés por la propuesta de INTELE.

3. MIRANDO AL FUTURO

Tras todo el trabajo realizado por la red INTELE y una vez conseguido el compromiso de las instituciones para la participación oficial de España en CLARIN y DARIAH, es hora de mirar al futuro y ponernos manos a la obra en la construcción de la nueva infraestructura para el castellano y el resto de lenguas cooficiales (euskera, gallego y catalán).

La futura infraestructura CLARIAH-es pretende subsanar estas carencias y cubrir las necesidades de los investigadores españoles. Al igual que en otros países europeos, CLARIAH-es será un consorcio de participación conjunta de España en las infraestructuras independientes CLARIN y DARIAH. Estará conformado por diversos nodos (véase figura 1) que deberán cumplir con un conjunto de requisitos mínimos para poder integrarse en él. Así, se requiere una permanencia en el consorcio de, al menos, cinco años, así como el compromiso de participar en él de forma coordinada con los otros nodos participantes y de proporcionar recursos para el desarrollo de la infraestructura. Mientras los nodos cumplan dichos requisitos previos, serán libres de buscar campos individuales de experiencia e interés. El respectivo trabajo de cada uno de los nodos participantes en el consorcio en diversas áreas del conocimiento dará cuerpo a la infraestructura compartida de CLARIAH-es. De esta forma, CLARIAH-es ofrecerá acceso a

investigaciones, herramientas, recursos y servicios especializados para todo tipo de temas relacionados con las Humanidades y las Ciencias sociales. Y, en la medida en que la colaboración entre nodos (ver Figura 1) será prioritaria, se facilitará considerablemente su participación conjunta en proyectos europeos, nacionales y locales, para con ello avanzar de forma efectiva en la transformación digital de la investigación española en las Humanidades y Ciencias Sociales, así como en su posicionamiento estratégico en proyectos y programas nacionales e internacionales, fundamentalmente en el contexto del Espacio Europeo de Investigación. Además, los nodos CLARIAH-es tendrán acceso al catálogo CLARIN, que incluye corpus, herramientas y servicios para la investigación y el análisis digital. A su vez, cabe destacar que contribuirán a CLARIN y DARIAH creando y ofreciendo herramientas y recursos para la investigación relacionada a nivel iberoamericano, incluida la investigación en castellano y lenguas cooficiales. A modo de ejemplo, cabe mencionar que el nodo vasco, actualmente en construcción, proporcionará herramientas para la investigación digital en una amplia gama de áreas, incluida la historia, la educación y la lengua durante los próximos cinco años.

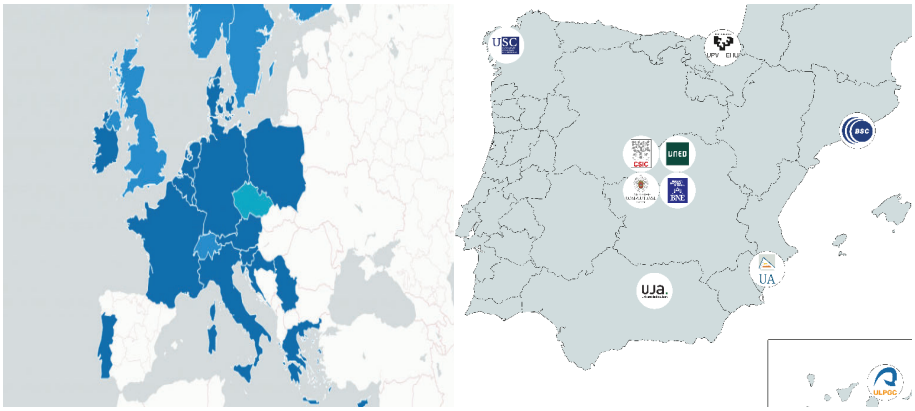

Figura 1 : El antes (izquierda) donde España no participa en ninguna de las infraestructuras y después (derecha) nodos comprometidos con el consorcio CLARIAH-ES

Aunque por el momento no se ha definido en todos sus aspectos, se ha presentado una primera propuesta de modelo de gobernanza para la futura infraestructura CLARIAH-es, para cuyo diseño se ha tomado como modelo el de su homólogo alemán. Dicho modelo, destinado a ayudar a

coordinar los distintos nodos y a identificar su lugar dentro de CLARIN y DARIAH, se erigiría en torno a seis paquetes de trabajo : 1) Datos de investigación, estándares y procedimientos; 2) Herramientas y entornos virtuales de investigación; 3) Participación de la comunidad, divulgación/difusión; 4) Capacitación y promoción de habilidades de investigadores jóvenes; 5) Integración técnica y coordinación de desarrollos técnicos; e 6) Infraestructura y administración. Esta distribución del trabajo a acometer pone de manifiesto que CLARIAH-es será una infraestructura multifacética e integral, que no se centrará únicamente en un aspecto tecnológico, como el alojamiento de datos o el desarrollo de software, sino que, además, creará recursos, organizará eventos y ofrecerá apoyo, servicios y consultoría a investigadores y grupos de investigación. Así, CLARIAH-es proporcionará datos que podrán utilizarse de manera sencilla para una variedad de propósitos con poca o ninguna necesidad previa de conocimientos técnicos. Pero también ofrecerá capacitación para desarrollar recursos y tecnología para la infraestructura, analizará lo que aún se requiere para garantizar que estas tecnologías se puedan usar de manera efectiva para la investigación y difundirá las actividades actuales en las que está involucrada la infraestructura. Es importante destacar que todas las herramientas y servicios dentro de la infraestructura serán evaluados y monitoreados, lo que permitirá hacer un seguimiento de los recursos que faltan o están en desarrollo y ayudará a determinar si existen herramientas con aplicaciones similares o si algunas herramientas son más adecuadas para ciertas tareas que otras.

3.1. Propuesta de red estratégica CLARIAH-ES

Para contribuir a afianzar el proceso de creación de la infraestructura CLARIAH-es, varias instituciones y futuros nodos previamente involucrados en INTELE han conseguido articular una propuesta común para una nueva red estratégica, también denominada CLARIAH-ES, que vertebrará y respaldará la adhesión oficial de España en CLARIN y DARIAH desde el próximo año 2023. El objetivo de dicha propuesta es auxiliar en el diseño y despliegue de la infraestructura CLARIAH-ES y de su relación con CLARIN y DARIAH, así como en la coordinación y gestión de sus primeros años de funcionamiento. Esto conlleva, entre otras tareas, identificar las necesidades de investigadores, grupos y proyectos que necesitan el apoyo de contenidos,

herramientas y recursos digitales; promover y facilitar la utilización de las infraestructuras, tanto propias, como de CLARIN y DARIAH, así como la participación en ellas, por investigadores españoles y comunidades de investigación que están trabajando en las Humanidades, Artes y Ciencias Sociales; y desarrollar y difundir las actividades de los centros participantes en la red y en las infraestructuras CLARIN y DARIAH. CLARIAH-ES también deberá participar en los distintos comités y grupos de trabajo de ambas infraestructuras, así como en sus reuniones, talleres y conferencias anuales, lo que facilitará la realización de informes sobre el progreso en el diseño de la infraestructura y la evaluación de ésta, otro de los fines de la red. En el marco de dicha participación, la red se compromete a organizar con carácter anual una serie de talleres sobre la investigación que es posible desarrollar gracias a los servicios, recursos y herramientas de las infraestructuras de investigación en Humanidades, Artes y Ciencias Sociales. Estos talleres serán un espacio indispensable para que la comunidad pueda conocer las posibilidades que ofrecen las infraestructuras CLARIN y DARIAH, y un foro que permitirá analizar las necesidades de CLARIAH-ES, establecer conexiones/contactos y colaboraciones entre los grupos de investigación, publicar los trabajos y proyectos más relevantes de la comunidad y colaborar en la búsqueda de soluciones a los problemas abiertos de forma conjunta.

Los ambiciosos objetivos de CLARIAH-ES solo se pueden lograr reuniendo los recursos necesarios en términos de datos, instalaciones informáticas y conocimiento que no están al alcance de ningún grupo de investigación en España. Por esta razón, la red estratégica CLARIAH-ES reúne a investigadores de diez centros de investigación líderes en Tecnologías del Lenguaje, Inteligencia Artificial, Computación de Alto Rendimiento, Humanidades y Ciencias Sociales, así como al principal centro informativo y documental sobre la cultura escrita, gráfica y audiovisual española e iberoamericana. A su vez, la red está formada por un grupo multidisciplinar de expertos informáticos, expertos lingüísticos en castellano y las lenguas cooficiales del estado (catalán, euskera, y gallego) y expertos en transición digital en las áreas de Humanidades, Artes y Ciencias Sociales. Cada uno de estos grupos aporta piezas clave de tecnología, conocimientos y datos necesarios que, juntos, forman un laboratorio de investigación distribuido que reúne la experiencia complementaria requerida para desarrollar correctamente esta iniciativa. El *know-how* combinado de los referidos grupos, que están a la vanguardia de

todas las técnicas que se aplicarán en CLARIAH-ES, y su participación con anterioridad en proyectos y publicaciones relacionados, garantizan que se puedan cumplir los objetivos de la red.

Los participantes en la red, entre los que se encuentran el centro Spanish CLARIN-K, el centro IMPACT CLARIN-K y un Cooperating Partner de DARIAH, están trabajando con CLARIN y DARIAH desde hace años y cuentan con el apoyo de la CRUE, el CSIC, la RAE, la BNE, el Instituto Cervantes y otros institutos, redes de investigación y asociaciones científicas. Cabe recordar que los diez grupos que actualmente conforman el consorcio CLARIAH-ES son la UPV/EHU (HiTZ), la Universidad de Santiago de Compostela (Instituto da Lingua Galega y CiTIUS), la Universidad de Alicante (Biblioteca Virtual Cervantes), la UNED (LINDH), el BSC, la Universidad Complutense de Madrid (UCM), la Universidad de Jaén (CEATIC), la ULPGC (IATEXT), el CSIC (Centro de Ciencias Humanas y Sociales) y la Biblioteca Nacional de España (BNE). Estos primeros nodos y los futuros de CLARIAH-ES cuentan con recursos propios de las universidades y con el apoyo de las comunidades autónomas (País Vasco, Galicia, Valencia, Madrid y Canarias, por citar algunas) para cubrir la aportación en especie (*in-kind*) necesaria para participar en CLARIN y DARIAH (alrededor de 1.5M€/año durante cinco años) con un compromiso de aportar el trabajo de 24 personas FTE (*full time equivalent*).[19] Garantizadas estas aportaciones mínimas durante los cinco primeros años, cualquier investigador del estado tendrá acceso desde su ordenador a ambas infraestructuras y a todos los datos, la tecnología, infraestructura y recursos necesarios para investigar en la nube con las herramientas digitales y de inteligencia artificial más avanzadas.

Formando parte de CLARIN y DARIAH, los investigadores tendrán acceso a instalaciones europeas de investigación únicas que incluyen estándares, datos, herramientas, computación, métodos, comunidades de interés, iniciativas o proyectos europeos. Como consecuencia, esperamos un incremento de proyectos de investigación y de la producción científica, así como una mejora del posicionamiento nacional e internacional de los investigadores en las áreas de las Humanidades, Artes y Ciencias Sociales. Consideramos, también, que la red estratégica CLARIAH-ES ofrecerá, entre otros, los siguientes resultados :

19 Cabe mencionar que la cuota (*fee*) de ambas infraestructuras, unos 200K€/año durante cinco años, la aportará el Ministerio de Ciencia e Innovación.

- Una mayor visibilidad de la nueva infraestructura CLARIAH-ES.
- Una mayor sostenibilidad y visibilidad de los resultados de proyectos de investigación nacionales.
- Una mayor visibilidad de la investigación nacional en las eHumanidades a nivel europeo con las ERIC CLARIN y DARIAH y el resto de infraestructuras europeas que participan en la iniciativa SSHOC[20].
- Un incremento de las oportunidades de investigación y colaboración a nivel nacional, europeo e iberoamericano.
- Un aumento de la interdisciplinariedad y multidisciplinariedad.
- Un aumento de las oportunidades de financiación y de consecución de proyectos a nivel europeo a través de la inserción, fruto de la colaboración en esta red, de España dentro de las infraestructuras.
- Una mayor interacción con la industria cultural y creativa (GLAM, en sus siglas en inglés).
- Una mayor interacción con los agentes del Perte de la Nueva Economía de la Lengua.

En síntesis, la red estratégica CLARIAH-ES espera un impacto real en la sociedad por cuanto permitirá un mayor intercambio de conocimiento, datos, tecnologías, infraestructuras, habilidades, experiencia, oportunidades y buenas prácticas, lo que incrementará el potencial de los resultados de los proyectos de investigación. A su vez, contribuirá a garantizar la sostenibilidad de herramientas y servicios para la investigación, favorecerá la generación de entornos colaborativos de trabajo y contribuirá a aumentar su influencia a nivel europeo, iberoamericano e internacional, así como a aumentar las oportunidades de financiación de las infraestructuras de investigación. Y, por último, contribuirá al desarrollo del sector productivo de la industria cultural y creativa.

20 https://sshopencloud.eu/

REFERENCIAS

BEL, Núria, GONZÁLEZ-BLANCO, Elena. e IRUSKIETA, Mikel. CLARIN Centro-K-español. *Procesamiento del Lenguaje Natural*, 2016, vol. 57, p. 151-154.

GIETZ, P. *et al*. TextGrid and eHumanities. In *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing E-SCIENCE*, 2006, IEEE Computer Society. Amsterdam.

IRUSKIETA, Mikel. y BEL, Núria. CLARIN-K Centre Spain : una infraestructura orientada al usuario, 2017. LINHD-UNED. Escuela de Verano Humanidades Digitales.

KALTENBRUNNER, Wolfgang. Digital Infrastructure for the Humanities in Europe and the US : Governing Scholarship through Coordinated Tool Development, *Computer Supported Cooperative Work (CSCW)*, 2017, vol. 26, p. 275–308. [DOI 10.1007/s10606-017-9272-2].

KRAUWER, S., y HINRICHS, E. The CLARIN research infrastructure : resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluatio*n, 2014, p. 1525-1531. European Language Resources Association (ELRA).

ROMARY, L. DARIAH-Shaping European Digital Research in the Arts and Humanities. *ERIC DARIAH, l'alliance des humanités et du numérique*, nov. 2014, Paris, France.

STEURS, F. Clarin, Clariah and Dariah : towards a full infrastructure for digital humanities in Europe. In *LexMC : Lexical Data Masterclass*, dic. 2017. Berlin, Germany.

KALMAN, T. y WANDL-VOGT, E. DARIAH-ERIC Towards a sustainable social and technical European eResearch Infrastructure for the Arts and Humanities. *e-IRG Workshop*, nov. 2014, Rome, Italy. (hal-01081479)

VÁRADI, T. *et al*. CLARIN : Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC)*, 2008, Marrakech, Marruecos, European Language Resources Association (ELRA).

Boletín ANABAD. LXXII (2022), NÚM. 2, ABRIL-JUNIO. MADRID. ISSN: 2794-0519 (USB) - 2444-7293 (Internet)

"Yo lo que en realidad necesito es…" User Stories para la exploración de necesidades sobre infraestructuras CLARIAH-DE y Text+

José Calvo Tello, Nanette Rissler-Pipka

José Calvo Tello, *State and University Library Göttingen,*
https://orcid.org/0000-0002-1129-5604,
Nanette Rißler-Pipka, *Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG),*
https://orcid.org/0000-0002-0719-9003,
Dirección de correspondencia: calvotello@sub.uni-goettingen.de

Palabras clave: Historias de usuario, infraestructuras, humanidades digitales

Keywords: User stories, infrastructures, digital humanities

Resumen:

Cualquier proyecto o institución de infraestructura científica debe tomar una serie de decisiones sobre qué soluciones ofrece, por ejemplo, a qué comunidades se dirige, qué formatos y soluciones informáticas aplica, qué vocabularios utiliza para su descripción o qué preguntas de investigación consideran relevante. En esta presentación, se explicó cómo se buscaron esas respuestas mediante la integración de los usuarios. Específicamente, los usuarios podían enviar historias de usuarios (en inglés, *user stories*) en las que describen su

situación actual, los retos a los que se enfrentan, posibles soluciones y su disponibilidad a acompañar el proceso de revisión e implementación. La acción de recopilación de historias de usuarios tuvo lugar en el marco de la solicitud de consorcios para el nuevo programa National Research Data Infrastructure Germany (NFDI), en concreto del consorcio Text+, cuyo principal objetivo es encargarse de datos textuales, en primer lugar para las disciplinas de lingüística y estudios literarios. El consorcio se estructura principalmente mediante un áreas de infraestructura y operaciones y tres tipos de datos (*data domains*): colecciones, recursos léxicos y ediciones. Entre las instituciones que integran el consorcio, se encuentran la biblioteca nacional alemana y la biblioteca estatal y universitaria de Göttingen como instituciones solicitantes, así como la biblioteca Herzog August Wolfenbüttel (HAB) y el Archivo Literario Alemán (Deutsches Literaturarchiv - DLA) como instituciones participantes.

La respuesta de la comunidad a esta petición de contribuciones fue más positivo de lo esperado, con 120 historias de usuarios de un total de 163 de investigadoras e investigadores diferentes. La página web del consorcio ha publicado las historias de usuarios, además de sus metadatos y un informe describiendo el proceso y su análisis. Para este análisis, cada historia fue anotada con diferentes tipos de categorías como la disciplina, qué criterios FAIR con especialmente importantes, así como un conjunto de categorías que describen la situación y las necesidades de la usuaria o usuario. Estas categorías nos permitieron visualizar las necesidades de cada una de las disciplinas y de los tipos de datos. En la actualidad, el consorcio utiliza estos datos para tomar decisiones sobre los siguientes pasos, además de planear una segunda llamada de historias de usuarios.

Abstract:

In research data infrastructure projects or consortia, the decision making process regarding offers for dedicated communities needs to rely on communication. What file formats, standards and computational solutions does the community work with? What are the vocabularies they use to describe their data? What are the most important research questions? In this presentation we explain how to get the desired answers by integrating the user into the decision making process. We asked the community to tell their stories and describe the situation, the challenges they are facing, if they have ideas for solutions and if they would be ready to be further part of the implementation and

review process. The user story call took place in the context of the National Research Data Infrastructure (NFDI) in Germany, in particular the Text+ consortium which is dedicated to the text- and language-based research data and thus particularly to disciplines like linguistics and literary studies. The consortium is structured around its task areas Infrastructure/Operations and the data domains: collections, lexical resources and editions. Among others, the German National Library and the Göttingen State and University Library are applicant institutions and the Herzog August Bibliothek Wolfenbüttel (HAB) and the German Literature Archive (Deutsches Literaturarchiv – DLA) are participant institutions.

The feedback by the community to the call for user stories was much better than could be hoped for: we received 120 user stories from 163 individual researchers. On the Text+ website we published all the stories together with metadata and information about the process of analysis. To prepare the analysis, every story was annotated according to three different types of categories: discipline, FAIR principles, and categories for the particular situation and requirements. With the help of these categories, we were able to create visualisations for each discipline and data domain. Today, the consortium uses the results of the user stories and the data created through the analysis for decision making and plans a second round of the user story call.

REFERENCES

BARBOT, Laure et al., 2020. *SSH Open Marketplace. Public Consultation for the DARIAH Community (Version 1.0)* [online]. Zenodo. DOI 10.5281/zenodo.3935345.

BARBOT, Laure et al., 2019. *SSHOC D7.1 System Specification - SSH Open Marketplace (Version 1.0)* [online]. Zenodo. DOI 10.5281/zenodo.3547648.

BIERWIRTH, Maik, et al., 2020. *Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung (Version 1.0)* [online]. Zenodo. DOI 10.5281/zenodo.3895209.

BRÜNGER-WEILANDT, Sabine, et al., 2020. *Memorandum of Understanding by NFDI Initiatives from the Humanities and Cultural Studies. (Version 2.0)* [online]. Zenodo. DOI 10.5281/zenodo.4045000.

BUDDENBOHN, Stefan, BARTHAUSER, Raisa and DUNG, Joseph, 2020. *User Requirements for the SSH Open Marketplace - Results from a SSHOC Workshop on January 30th* [online]. Zenodo. DOI 10.5281/zenodo.3759174.

CALVO TELLO, José, et al., 2021. *Text+ User Stories Data*. DARIAH-DE Repository [online]. DOI http://dx.doi.org/10.20375/0000-000E-67ED-4.

GEHRING, Petra, 2020. *Lernen aus der ersten Auswahlrunde. 2. NFDI-Konferenz 8./9. Juli 2020* [online]. Available in: https://www.dfg.de/download/pdf/foerderung/programme/nfdi/nfdi_konferenz_2020/vortrag_gehring.pdf.

German Council for Scientific Information Infrastruktures (RfII), 2016. *Enhancing Research Data Management: Performance through Diversity. Recommendations regarding structures, processes, and financing for research data management in Germany* [online]. Göttingen. Available in: https://rfii.de/?p=2075.

German Council for Scientific Information Infrastructures (RfII), 2018. *Wide Impact for Research: NFDI Consortia as Stakeholders. Third discussion paper on the development of a national research data infrastructure (NFDI) in Germany* [online]. Göttingen. Available in: https://rfii.de/?p=3818.

GLÖCKNER, Frank Oliver, et al., 2019. *Berlin Declaration on NFDI Cross-Cutting Topics (Version 1.0)* [online]. Zenodo. Available in: 10.5281/zenodo.3457213.

NFDI Expert Committee, 2020. *The development of the National Research Data Infrastructure (NFDI). Second statement of the NFDI Expert Committee* [online]. Available in: https://www.dfg.de/download/pdf/foerderung/programme/nfdi/stellungnahme_nfdi_201112_en.pdf.

RISSLER-PIPKA, Nanette et al., 2021. *Community Involvement in Research Infrastructures: The User Story Call for Text+* [online]. Zenodo. DOI https://doi.org/10.5281/zenodo.5384085.

SCHILL, Kerstin, 2020. *Der Aufbau der NFDI. 2. NFDI-Konferenz 8./9. Juli 2020* [online]. Available in: https://www.dfg.de/download/pdf/foerderung/programme/nfdi/nfdi_konferenz_2020/vortrag_schill.pdf.

TEXT+ CONSORTIUM, 2021. *User stories* [online]. Available in: https://www.text-plus.org/forschungsdaten/user-stories/.

WARWICK, Claire, 2012. Studying users in digital humanities. In WARWICK, Claire, TERRAS, Melissa, and NYHAN, Julianne (editors), *Digital Humanities in Practice* [online]. Facet, pp. 1-22. DOI doi:10.29085/9781856049054.002.

WILKINSON, Marc D., et al., 2016. *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data, 3, 160018. ISSN 2052-4463. DOI https://doi.org/10.1038/sdata.2016.18.