Modelo de Gobernanza de las Infraestructuras Lingüísticas

Plan de Impulso de las Tecnologías del Lenguaje

Núria Bel

12/2016

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para la Sociedad de la Información y la Agenda Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

ÍNDICE

1 INTRODUCCIÓN 6

2 ¿QUÉ DEBE SER GOBERNADO? 11

2.1 OPERACIONES 12

2.2 PLANIFICACIÓN ESTRATÉGICA 13

3 ¿QUIÉN DEBE PARTICIPAR EN LA GOBERNANZA DE LAS INFRAESTRUCTURAS? 14

3.1 PROVEEDORES DE DATOS Y PROCESADORES 15

3.1.1 UNIVERSIDADES Y ORGANISMOS PÚBLICOS DE INVESTIGACIÓN 16

3.1.2 ORGANISMOS DE NORMALIZACIÓN LINGÜÍSTICA 17

3.1.3 REPOSITORIO DE DATOS 18

3.1.4 UNIVERSIDADES Y CENTROS DE INVESTIGACIÓN 18

3.2 GRUPOS DE ELABORACIÓN DE ESTÁNDARES 19

3.3 AGENCIAS DE FINANCIACIÓN 19

3.4 USUARIOS 20

4 Antecedentes e iniciativas relacionadas 21

4.1 CLARIN 21

4.1.1 CLARIN EN LOS ESTADOS MIEMBROS 24

4.2 META-SHARE 25

4.3 ELRA/ELDA 27

4.4 LDC 30

4.5 CREL 31

4.6 IHTSDO 32

4.7 BSC-CNS 33

4.8 IRTA 35

4.9 PROYECTO VISC+ 36

5 ESTÁNDARES EN LAS TECNOLOGÍAS DEL LENGUAJE 37

6 CUESTIONES A TENER EN CUENTA PARA LA GOBERNANZA 41

6.1 Cuestiones en la Planificación estratégica 43
6.2 Cuestiones en las Operaciones 47
7 RECOMENDACIONES 48
7.1 Principios subyacentes al modelo de gobernanza 49
7.2 Planificación. Órganos de la gobernanza 50
7.3 Toma de decisiones 52
7.4 Comisiones delegadas 53
7.4.1 Comisión de planificación estratégica 54
7.4.2 Comisión de evaluación, adquisición y nuevos desarrollos 54
7.4.3 Comisión de estándares y licencias 55
7.4.4 Comisión de servicios 55
7.5 Operaciones. Modelos y organización 55

ÍNDICE DE FIGURAS

Figura 1: Una infraestructura ha de servir para mejorar procesos de terceros y ha de garantizar acceso e interoperabilidad. 6
Figura 2: Captura de pantalla de la página que da acceso a los servicios ofrecidos por CLARIN (acceso 22-11-2016) 23
Figura 3: Diagrama de la relación entre los órganos de planificación y operaciones del ERIC CLARIN 24
Figura 4: Diagrama de la relación entre los órganos de planificación y operaciones de la European Language Resources Association 30
Figura 5: Diagrama de la relación entre los órganos de planificación y operaciones del BSC-CNS que figura en su página web 34
Figura 6: Diagrama de la relación entre los órganos de planificación y operaciones del Institut de Recerca en Tecnologia dels Aliments 36
Figura 7: Resultado del análisis morfosintáctico de FreeLing en formato XML 38
Figura 8: Resultado del análisis morfosintáctico de IxaPipes en formato XML 38
Figura 9: Resultado del análisis morfosintáctico de Stanford CoreNLP en formato XML 39
Figura 10: Diagrama de la relación entre los órganos que participan en la planificación de las infraestructuras lingüísticas del Plan TL. Recomendaciones. 53

Figura 11: Diagrama de la relación entre los órganos que participan en las operaciones de las infraestructuras lingüísticas del Plan TL. Recomendaciones..... 55

ÍNDICE DE TABLAS

Tabla 1: Gestión de estándares y otras características de las infraestructuras lingüísticas de los antecedentes..... 41

Tabla 2: Características de las infraestructuras lingüísticas de los antecedentes y modelo de negocio 42

Modelo de Gobernanza de las Infraestructuras Lingüísticas

1 INTRODUCCIÓN

El objetivo de este documento es ser la base para una discusión sobre la gobernanza de las infraestructuras lingüísticas propuestas en el Eje 1 del Plan de Impulso de las Tecnologías del Lenguaje (Plan TL). Este informe se entregará a la Secretaria de Estado para la Sociedad de la Información y la Agenda Digital (SESIAD) del Ministerio de Energía, Turismo y Agenda Digital.

Figura 1: Una infraestructura ha de servir para mejorar procesos de terceros y ha de garantizar acceso e interoperabilidad.

Las infraestructuras lingüísticas consisten en procesadores del lenguaje natural, datos lingüísticos y herramientas de evaluación para procesadores y datos. En referencia a las infraestructuras lingüísticas, hablaremos de procesadores para referirnos a herramientas lingüísticas: aquellas que inducen información sobre las palabras, o las combinaciones de palabras en uno o más textos. En cuanto a los datos, que sirven principalmente para entrenar procesadores, son colecciones de textos, colecciones de textos enriquecidos (es decir, anotados con información explícita) y recursos léxicos (por ejemplo, diccionarios bilingües, listas de terminología con información lingüística, vocabularios o diccionarios con información semántica). Por evaluación nos referimos a la tarea de medir la precisión de los procesadores: su actuación con respecto a su objetivo y a la de medir la calidad de los datos, especialmente en lo referente a la anotación (la consistencia y estandarización de la misma como factores más importantes). El Informe sobre el estado de las tecnologías del lenguaje en España dentro

de la Agenda Digital[1] hace una descripción de los tres componentes y de los recursos existentes ya para el español y las lenguas co-oficiales.

El conjunto de estos elementos se denomina también infraestructura y en este informe utilizaremos este nombre para referirnos al conjunto y, por extensión, a la organización que provee este conjunto de recursos.

Las infraestructuras lingüísticas del Eje 1 del Plan TL pretenden ser el motor de una substitución del modelo actual en el que cada desarrollador (empresa y academia) debe construir su propia infraestructura –procesadores y datos, con la consiguiente reduplicación de esfuerzos y pérdida de tiempo, por un modelo en el que un catálogo-repositorio de procesadores y datos disponibles permita satisfacer rápidamente diferentes niveles de necesidad para agilizar la creación de productos: desde recursos limitados pero gratuitos para el desarrollo de prototipos y demostradores, hasta datos y procesadores que puedan ser utilizados con licencias comerciales para productos finales. Para conseguir el cambio de modelo, además de identificar la existencia de procesadores y datos, hay que asegurar su accesibilidad, eliminando las barreras técnicas, legales y humanas que limitan actualmente la reutilización de datos y procesadores.

Desde el punto de vista técnico, el objetivo de la reutilización de datos y procesadores obliga a acometer la selección de unas especificaciones técnicas que la garanticen: los datos y procesadores han de tener unas características técnicas, descritas y conocidas a priori. La definición consensuada de estas especificaciones y su aplicación real han de convertirlas en estándares para que proveedores y desarrolladores maximicen el potencial de reutilización al crear componentes y datos interoperables. La utilización de estándares facilita el desarrollo de productos puesto que no es necesario invertir en la adaptación de herramientas para cubrir diferentes formatos por ejemplo, cuando se plantean herramientas multilingües o cuando se substituye un procesador por otro. Por último, la utilización de estándares también facilita la evaluación de datos y procesadores. El modelo de gobernanza de las infraestructuras lingüísticas tiene que prever la estructura de gobierno y los mecanismos de decisión que aseguren el consenso en la elección de estándares y que garanticen su aplicación para crear un entorno de interoperabilidad.

[1] *http://www.agendadigital.gob.es/planes-actuaciones/tecnologias-lenguaje/Bibliotecaimpulsotecnologiaslenguaje/Material%20complementario/Informe-Tecnologias-Lenguaje-Espana.pdf , Capítulo 5.*

Para hacer posible la reutilización de procesadores y, muy especialmente, de datos, ha de abordarse también la selección de un modelo de licencias que garantice los derechos legales de productores y usuarios. La reutilización de datos que son textos escritos y que se refieren a terceros suscita por un lado cuestiones de copyright que están siendo debatidas actualmente[2], y por otro de tratamiento de datos privados. Ambos merecen una atención especial para el desarrollo y el gobierno de la infraestructura. El modelo de gobernanza ha de asegurar el respeto a las restricciones legales y la capacidad de negociación con terceros sobre los derechos de propiedad y de copia de los procesadores y datos de las infraestructuras. Las infraestructuras también han de poder ofrecer servicios, como la anonimización de datos, para asegurar a los proveedores de datos que no habrá infracciones en cuanto al tratamiento de datos privados.

Así pues, en el Plan TL la finalidad de desarrollar infraestructuras lingüísticas es ofrecer un número de procesadores, conjuntos de datos y sistemas de evaluación que puedan ser reutilizados por la industria y la academia para la creación de nuevas herramientas o de productos que requieren procesamiento lingüístico. El Plan TL, mediante la creación de las infraestructuras, asume la misión fomentar la innovación y la transferencia de tecnología y conocimiento: ofrecer tecnología para la innovación, reducir la inversión y el tiempo necesario para la puesta de productos en el mercado.

Las infraestructuras se definen así por el objetivo de ser un bien colectivo en el que se aportan bienes de diferentes proveedores para ser utilizados por terceros. Por ello, en la definición de la gobernanza se deberá tener en cuenta a todos los sectores implicados: proveedores de datos, desarrolladores de procesadores, usuarios-beneficiarios de la infraestructura y, también, la agencia financiadora al menos en el estadio de construcción previsto por el Plan TL ya que este plan pretende lograr una serie de objetivos que la agencia que lo finanza ha de poder priorizar así como controlar su cumplimiento. La gobernanza de las infraestructuras ha de dotarlas de mecanismos de interacción entre todos estos agentes para asegurar una planificación estratégica que garantice su puesta en marcha, despliegue óptimo y, muy especialmente, su sostenibilidad. Pero, por otro lado, ha de dotarla de suficiente autonomía para hacerla eficiente en las operaciones de gestión.

[2] *Ver directivas sobre el copyright en la UE: https://ec.europa.eu/digital-single-market/en/copyright, en particular la nueva propuesta (14-09-2016) sobre "Text and data mining": http://europa.eu/rapid/press-release_MEMO-16-3011_en.htm que, no obstante, se refiere únicamente a la investigación.*

La creación de infraestructuras para fomentar la innovación tiene ya algunos antecedentes y también existen varias iniciativas para la creación específicamente de infraestructuras lingüísticas en Europa. En este documento se ha hecho una revisión de estos antecedentes en el ámbito y se describe su modelo de gobernanza para ofrecer posibles modelos para las nuevas infraestructuras lingüísticas del Plan TL. Por ejemplo, la gran mayoría tiene modelos de gobernanza que separa la planificación estratégica de la operativa y esta separación ya se tiene en cuenta desde el principio en este informe. El modus operandi de las infraestructuras del Plan TL está por concretar y su definición será parte de las tareas iniciales y prioritarias que los órganos de gobernanza de las mismas habrán de acometer. No obstante, cabe pensar que no diferirá radicalmente del funcionamiento de estas otras infraestructuras ya existentes, lingüísticas o de otros ámbitos (de los que se ofrece también una selección ilustrativa), cuyo objetivo es poner a disposición de terceros datos, herramientas y servicios con el objetivo de fomentar la capacidad de innovación de las empresas del sector al que se dirigen.

Además, en lo que se refiere a otras iniciativas específicas de infraestructuras lingüísticas, la lista de las mismas y su naturaleza y actividades permitirá también revisar con cuales de ellas se debería establecer contactos que permitan interactuar con las mismas. Esta interacción con otras infraestructuras también debería quedar garantizada por el modelo de gobernanza elegido.

El Plan TL contempla la creación de infraestructuras lingüísticas para las lenguas de España y se refiere, por tanto, al español, catalán, euskara y gallego. Procesadores y datos lo son para una lengua determinada y, por tanto, se pretende el desarrollo de infraestructuras que den la misma cobertura para cada una de nuestras lenguas. *El Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital*[3] hacía una revisión exhaustiva sobre la disponibilidad de datos, procesadores y herramientas de evaluación y analiza las necesidades que deberían ser cubiertas para igualar lo disponible para el inglés. En cuanto a la magnitud, el Plan TL se marca como objetivo el cubrir estas cuatro lenguas ofreciendo los mismos recursos que existen para el inglés, en este sentido la lengua más cubierta del mundo (*META-NET White Papers, 2012*[4]). Pero hay que acotar la magnitud de la infraestructura para asegurar la viabilidad de la iniciativa, además de promover la colaboración, la reducción de duplicaciones, y asegurar sinergias para abarcar la mayor parte de los recursos que la

[3] *http://www.agendadigital.gob.es/planes-actuaciones/tecnologias-lenguaje/Bibliotecaimpulsotecnologiaslenguaje/Material%20complementario/Informe-Tecnologias-Lenguaje-Espana.pdf , Capítulo 5.*

[4] *http://www.meta-net.eu/whitepapers/overview*

industria ya está demandando. Otra característica de las infraestructuras lingüísticas es el enorme potencial de crecimiento que tienen: no puede establecerse solamente la creación de unos determinados recursos y hablar de sostenibilidad solamente como explotación. La sostenibilidad de las infraestructuras se refiere al mantenimiento y explotación de lo creado gracias al Plan TL y, también, a garantizar que siga creciendo en número de recursos y en la adaptación de los recursos existentes a nuevos usos y a nuevas tecnologías, usos lingüísticos (desde un nuevo dominio de aplicación a unos nuevos criterios de ortografía, por ejemplo).

La gobernanza también ha de tener en cuenta la visibilidad de la iniciativa. Ya existen iniciativas que parcialmente abordan alguna de las cuestiones que tratará la infraestructura del eje 1, el mayor problema todavía es que buena parte de los actores implicados las desconocen. Es de la mayor importancia que la gestión de la infraestructura garantice, y de forma dilatada en el tiempo, difusión de sus contenidos y sus objetivos.

Por último, el esquema de gobernanza, la estructura y organización (y participantes necesarios) de las infraestructuras han de garantizar la capacidad de acciones coordinadas con los agentes relevantes en la definición e implementación de las medidas concretas que establece el Eje I del Plan TL:

> *Medida I.1. Seleccionar normas técnicas de interoperabilidad, políticas de licencias y mecanismos de protección de datos adecuados para la generación de recursos lingüísticos.*
>
> *Medida I.2. Adquirir o desarrollar herramientas comunes para la generación y evaluación de infraestructuras lingüísticas.*
>
> *Medida I.3. Elaborar y ejecutar un plan de desarrollo de infraestructuras lingüísticas. Realizar un inventario de infraestructuras lingüísticas actualmente disponibles. Evaluar la evolución de la cantidad, calidad y disponibilidad de infraestructuras lingüísticas.*
>
> *Medida I.4. Facilitar el acceso público a las infraestructuras lingüísticas existentes.*

En lo que sigue se abordan primero las preguntas a hacerse sobre la gobernanza. En la sección 2 se plantea la pregunta ¿qué debe ser gobernado? Se aporta una lista de tareas, separadas, como ya se ha mencionado, en Planificación Estratégica y Operaciones. En la sección 3 se plantea ¿Quién debe participar en la gobernanza? Se aporta una lista, pretendidamente exhaustiva, de los agentes relacionados con el ámbito. En la sección 4, se aportan detalles de varias iniciativas específicas de creación de infraestructuras lingüísticas o similares. En cuanto a las consideradas infraestructuras lingüísticas actualmente, la mayoría de ellas están activas: CLARIN, META-SHARE, ELRA/ELDA, en

Europa y LDC de Estados Unidos. Estas dos últimas son organizaciones distribuidoras de datos lingüísticos concebidas con la idea de mitigar el problema de la escasez de los mismos y, por tanto de su reutilización. La quinta, CREL ya no está operativa pero es el primer antecedente en España de una organización con el objetivo de constituir un conjunto de datos y procesadores. La Organización Internacional para el Desarrollo de Estándares de Terminología Sanitaria (IHTSDO) es una asociación que tiene como objetivo desarrollar, mantener y comercializar una terminología médica clínica (y programas de gestión relacionados) para ser utilizada en informática médica.

Además, se dan detalles de otras tres infraestructuras españolas (dos ya consolidadas y muy activas y una tercera en preparación) para otros dominios que no están relacionados con las lingüísticas del Plan TL, pero que comparten con éstas del Plan TL algunos aspectos: dar servicios y/o datos con el objetivo de fomentar la innovación y la transferencia de tecnología. BSC-CNS, una infraestructura de supercomputación; IRTA, para fomentar la innovación en el área de Tecnologías de los Alimentos, y el muy reciente Proyecto VISC+ que pretendía poner a disposición de los investigadores datos clínicos pero que no se ha constituido todavía. En la sección 5 se hace un pequeño aparte sobre iniciativas de estandarización en el área de las tecnologías del lenguaje, con el objetivo de documentar la importancia de la toma de decisiones de las infraestructuras en ámbitos como el de los estándares que garantizarán la interoperabilidad de todos los elementos de las infraestructuras. En la sección 6, se aborda la pregunta de cómo la gobernanza (actores, cadenas de mando y procedimientos, coordinación entre ellos y con otras iniciativas) aseguraría la viabilidad del proyecto, su estabilidad y sostenibilidad (preservación y crecimiento). Por último, en la sección 7 se hace una propuesta de modelo de gobernanza, que pretende ser la base para la discusión y finalmente la implementación de la iniciativa propuesta por el Plan TL.

2 ¿QUÉ DEBE SER GOBERNADO?

Como se ha avanzado, las infraestructuras lingüísticas son un bien colectivo que ha de ser dispuesto como servicios para terceros. Esta misión implica una serie de operaciones que garanticen básicamente el ofrecer un número de procesadores y datos, debidamente evaluados e interoperables a los que la industria, en particular pymes y start-ups que no disponen de capacidad de inversión para desarrollar su propia infraestructura, pueda acudir para desarrollar nuevos o mejores productos. Además, la gestión de las infraestructuras convierte al organismo responsable de ellas en un interlocutor especializado para llevar a cabo tareas de asesoría y evaluación (acreditación) para otros organismos y para empresas. Para asegurar el logro de estos objetivos y de forma muy importante el éxito y la sostenibilidad de los resultados, la gobernanza de la infraestructura requiere de órganos de

planificación estratégica: dirección y planificación temporal y estratégica de medidas y actividades, monitorización y control.

Detallamos ahora las tareas a cubrir separándolas en operaciones y planificación estratégica.

2.1 OPERACIONES

Un listado, no exhaustivo, de operaciones que habrán de llevar a cabo las infraestructuras lingüísticas son las siguientes:

Distribución

1. Distribución de recursos (datos y procesadores): censo y catálogo accesible y con posibilidad de descarga de datos/procesadores, y aplicación de herramientas de monitorización de consumo (estadísticas) para analizar la oferta y demanda.
2. Operativa de evaluación y certificación de calidad y usabilidad.
3. Oferta de servicios a usuarios: diseño y usabilidad y actividad contractual derivada.
4. Almacenamiento de datos-procesadores: mecanismos apropiados de almacenamiento seguro y acceso en caso de necesidades de confidencialidad, integridad y disponibilidad.

Adquisición y producción

5. Censo de proveedores de recursos
6. Búsqueda activa de recursos y proveedores: aplicación de los criterios de selección, creación y mantenimiento (actualización) de datos
7. Organización y documentación –metadatos-- de los datos y procesadores siguiendo estándares del ámbito (internacionales) que permitan buscar y encontrar los datos y saber detalles de la creación como documentación necesaria para su reutilización.
8. Actividad contractual de la adquisición de recursos
9. Preparación de los recursos para su distribución/utilización (empaquetado y licencias)
10. Vigilancia y acciones derivadas de las cuestiones éticas y legales (datos personales, copyright y licencia de acceso y utilización de los datos)

11. Producción y mantenimiento de datos y procesadores

12. Establecimiento de criterios técnicos en la creación de nuevos datos o procesadores

Difusión y formación

13. Ejecución de las campañas de difusión y formación relacionadas con estándares y tecnología.

14. Seguimiento y comunicación del logro de objetivos.

2.2 PLANIFICACIÓN ESTRATÉGICA

En cuanto a la planificación estratégica, una lista no exhaustiva de las tareas a gestionar que agrupamos en torno a cuatro cuestiones básicas:

Planificación

1. Elaboración de un modelo de plan de gestión de datos lingüísticos, recomendaciones como parte de la propuesta de financiación que anticipe la gestión posterior y que incluya la viabilidad económica.

2. Elaboración de un plan estratégico 2020-2025: Modelo de negocio: amortización y retorno; mantenimiento, escalabilidad y sostenibilidad de las infraestructuras.

3. Establecer relaciones con infraestructuras, proyectos e iniciativas relacionadas (Datos abiertos, EUDAT, CLARIN) en Europa y en Iberoamérica.

Evaluación, adquisiciones y nuevos componentes

4. Planteamiento y criterios técnicos para definir y ejecutar proyectos de nuevos datos, procesadores, y/o evaluación, y de su mantenimiento y usabilidad.

5. Estudios de evolución y masa crítica de recursos, y número de usuarios-contribuyentes. Evolución por necesidades, por oferta. Flexibilidad y escalabilidad de las infraestructuras.

Estándares y licencias

6. Vigilancia y selección de estándares de interoperabilidad de las infraestructuras, participación en grupos de decisión de estándares europeos e internacionales y

garantizar que las especificaciones de los mismos cubren y se cumplen para español y lenguas co-oficiales.

7. Vigilancia y selección de esquema de licencias para la distribución y utilización de datos y procesadores.

8. Vigilancia y selección de mecanismos de protección de datos personales y garantías de reutilización de textos con respecto a los derechos de propiedad intelectual.

9. Servicios

10. Planificación de actividades para maximizar el alcance, visibilidad y plan de difusión de las infraestructuras lingüísticas. Fomentar la creación y sostenibilidad, herramientas y buenas prácticas.

11. Elaborar informes y recomendaciones para la dirección del Plan TL y para otras instancias de la Administración Pública.

12. Proponer procedimientos para la creación, utilización y re-utilización de recursos y procesadores en las actividades de la administración pública.

13. Promocionar el ámbito del procesamiento del lenguaje natural y las infraestructuras lingüísticas en los planes estratégicos de la administración pública en investigación e innovación y desarrollo.

Todas estas tareas están relacionadas con el análisis y el diagnóstico de cuestiones que implican recabar información y asegurar la participación de diversos agentes/actores en la selección de propuestas para el modus operandi de las infraestructuras.

3 ¿QUIÉN DEBE PARTICIPAR EN LA GOBERNANZA DE LAS INFRAESTRUCTURAS?

La gobernanza ha de materializar el equilibrio de intereses mediante un sistema de reglas que rijan las relaciones entre los diferentes actores que pueden estar relacionados y que analizamos a continuación.

Los productores de datos son los que recolectan y editan esos datos (los limpian, estructuran, organizan en bases de datos, etc.) y los procesan (manual o automáticamente) para añadir información explícita que haga posible procesarlos para crear nuevos procesadores y herramientas. Los desarrolladores de los procesadores reutilizables (programas informáticos basados en algoritmos normalmente ya ampliamente conocidos) son investigadores de universidades y desarrolladores industriales que pueden seguir modelos públicos o proponer nuevos algoritmos. La evaluación (manual o automática), las comparaciones entre los resultados de los procesadores (o de procesos manuales de anotación) con materiales preparados que revelan la calidad de los resultados, tiene como actores principales nuevamente investigadores académicos[5] que se encargan de preparar los materiales y llevar a cabo las comparaciones.

Pero hay otros agentes que pueden ser relevantes para el ámbito de las infraestructuras lingüísticas: los que disponen de datos, especialmente textos, y los gestionan como bibliotecas, archivos de datos, organismos cuya misión es la normalización lingüística, editoriales y publicaciones en general, por ejemplo.

También incluimos en la lista de participantes en la gobernanza a los grupos de elaboración de estándares, al menos una conexión debe quedar asegurada. Por último, la gobernanza debería tener en cuenta, si no es obligada, la participación de las agencias financiadoras y de los usuarios a los que van dirigidas.

Hacemos en las siguientes secciones un repaso de estos agentes, describiendo sus características. En la sección 6 valoramos en qué medida la participación de cada uno de ellos puede favorecer el éxito en la misión de la infraestructura.

3.1 PROVEEDORES DE DATOS Y PROCESADORES

Gran parte de las operaciones de las infraestructuras lingüísticas están relacionadas con la identificación y colección de datos y procesadores existentes. La relación con los proveedores es por tanto fundamental para asegurar la viabilidad de la infraestructura. En especial en lo relativo a los datos, que son en su mayoría textos, hace que la lista de proveedores sea amplia.

[5] *Aunque hay algunas evaluaciones más institucionalizadas, por ejemplo en Estados Unidos las del trec.nist.gov*

3.1.1 UNIVERSIDADES Y ORGANISMOS PÚBLICOS DE INVESTIGACIÓN

Los datos primarios (los textos ya en colecciones o no, ya con información de temática o no) son recolectados por investigadores que los organizan, limpian, enriquecen con información explícita adicional y los convierten en conjuntos de datos procesables. Suelen utilizar formatos y estándares de facto relacionados con procesadores concretos. Esta labor es principalmente realizada por equipos de investigadores de universidades y centros de investigación[6]. Los proveedores pueden serlo de una o más de una lengua.

Aunque las empresas del sector[7] son también productores de datos específicos para sus aplicaciones, las empresas no suelen compartir sus datos ya que se convierten en un activo competitivo[8], y hacen de los centros públicos los principales suministradores de datos. En las universidades y OPIs, la creación de conjuntos de datos suele estar financiada por agencias públicas (programas de investigación y transferencia de ámbito estatal o autonómico, Programa Marco de Investigación y Desarrollo de la Unión Europea).

Los desarrolladores de procesadores también han sido tradicionalmente los investigadores de universidades y centros públicos. Estos procesadores, o al menos de versiones iniciales, han sido transferidos a la industria y en algunos casos son software abierto y con licencias que permiten usos comerciales. Como en el caso de los recursos, algunas empresas del sector han desarrollado procesadores propios (normalmente las grandes corporaciones como Google, IBM, Microsoft, etc.) y licencian su utilización en productos comerciales, cada vez más como servicios a través de APIS[9].

Los productores universitarios están agrupados en asociaciones y redes, en España: Sociedad Española del Procesamiento del Lenguaje Natural[10], Red Temática en Tratamiento de la información TIMM[11],

[6] Por ejemplo, el Centro Ramón Piñeiro http://www.cirp.es/ o la Fundación Elhuyar https://www.elhuyar.eu
[7] Los datos para traducción automática son un caso aparte ya que se trata de textos traducidos cuyos productores son agencias de traducción, y donde hay un incipiente mercado. Ver TAUS www.taus.net
[8] Hay excepciones como Google ngrams: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
[9] https://cloud.google.com/natural-language/ o https://www.luis.ai/ son versiones betas recientemente publicadas.

[10] http://www.sepln.org/
[11] http://timm.ujaen.es/

Red de Excelencia en Recursos para las Tecnologías del Lenguaje, RETELE[12]. También existen asociaciones a nivel europeo: META-NET[13], ELRA[14], como se describirá después en la sección 4.

3.1.2 ORGANISMOS DE NORMALIZACIÓN LINGÜÍSTICA

Los organismos con misión de normalización lingüística son productores de datos, en mayor medida de diccionarios y terminología monolingüe y bilingüe para consumo humano, pero también y cada vez en mayor medida, para la producción de estos recursos tradicionales usan colecciones de textos recopilados por ellos mismos o por terceros, así como procesadores lingüísticos que facilitan la explotación de los datos. Cada lengua dispone de un organismo que recibe el encargo. Los que incumben al plan son los siguientes: para español, la Real Academia Española[15](RAE) y algunas academias de las reunidas en el Instituto de España: Real Academia de Ciencias[16] y Real Academia de Medicina[17]; para catalán, el Institut d'Estudis Catalans[18] y el TERMCAT[19] (Centre de Terminologia de Catalunya); para gallego, la Real Academia Galega[20] y el TERMIGAL (Servizo de Terminoloxía Galega)[21]; para euskara, la Real Academia de la Lengua Vasca[22] y UZEI[23], la asociación que mantiene, entre otros recursos, Euskalterm[24].

También incluiremos aquí a organismos de la administración pública que disponen de recursos terminológicos (también tesauros, ontologías y otros recursos semánticos) en áreas especializadas. Un ejemplo significativo es la terminología SNOMED-CT en español (suministrada por el Ministerio de Sanidad, Servicios Sociales e Igualdad)[25], en catalán (suministrada por TICSalud)[26] (también se está trabajando en la versión vasca que tiene prevista su publicación en 2017).

[12] http://retele.linkeddata.es
[13] http://www.meta-net.eu/
[14] http://www.elra.info/en/
[15] http://www.rae.es/
[16] http://www.rac.es/5/5_1.php
[17] *http://www.ranm.es/terminolog%C3%ADa-m%C3%A9dica.html*
[18] *http://www.iec.cat*
[19] *http://www.termcat.cat*
[20] *http://academia.gal/recursos*
[21] *https://www.cirp.es/w3/proxectos/proxecto-termigal.html*
[22] *http://www.euskaltzaindia.eus*
[23] *http://www.uzei.eus/*
[24] *http://www.euskara.euskadi.eus/r59738/es/contenidos/informacion/euskalterm/es_7553/euskalte rm.html*
[25] http://www.msssi.gob.es/profesionales/hcdsns/areaRecursosSem/home.htm
[26] http://www.ticsalut.cat/estandards/terminologia/recursos/

3.1.3 REPOSITORIO DE DATOS

En gran medida los datos de las infraestructuras lingüísticas son textos, grandes colecciones de textos. Por tanto, se convierten en potenciales proveedores aquellas entidades que archivan obras textuales.

3.1.4 UNIVERSIDADES Y CENTROS DE INVESTIGACIÓN

Las Universidades y centros de investigación, como organizaciones, son actualmente los principales repositorios para la custodia y preservación de los datos producidos por sus investigadores, muchos de ellos en forma de textos. Dentro de estos organismos, los servicios de informática, biblioteca, y servicios de asesoría sobre derechos, licencia y publicación suelen ofrecerse como infraestructura para la investigación, pero su papel también empieza a centrarse en el almacenamiento, preservación y difusión de la información para fomentar el acceso abierto y su reutilización[27].

3.1.4.1 Repositorios Institucionales y Centros de Datos

En algunos países europeos, ya existe la figura de Archivo de Datos, como el DANS[28] de Holanda. Su papel es el almacenamiento a corto y largo plazo de conjuntos de datos. Sus tareas se concentran en asegurar el uso de estándares de interoperabilidad entre repositorios, mantener la fiabilidad y robustez de los accesos y desarrollar mecanismos para la migración (y recuperación) de datos entre repositorios. Hay grandes organizaciones como EUDAT[29] que coordinan institucionalmente las políticas de almacenamiento de datos en los diferentes países de la Unión.

Hay que tener en cuenta también la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público (RISP) que fomenta la publicación de datos de la administración pública como datos abiertos a los que se accede desde portales específicos. Por ejemplo: http://datos.gob.es/ o https://data.europa.eu/euodp/en/data. También en el ámbito de las empresas se constata una sensibilización hacia la política de datos abiertos.

3.1.4.2 Bibliotecas

Después del esfuerzo en digitalización llevado a cabo, las bibliotecas también son repositorios de datos primarios para las tecnologías del lenguaje que pueden estar sujetos a diferentes niveles de protección de copyright.

---

[27] *Los proyectos openAire y openMinTED, por ejemplo: (https://www.openaire.eu/ y http://openminted.eu/*
[28] *https://dans.knaw.nl/en*
[29] *https://www.eudat.eu/*

3.1.4.3 Publicaciones y Editoriales

También como consecuencia de la digitalización en los procesos, se han convertido en productores y almacenes de datos primarios (textos) que pueden estar sujetos a diferentes niveles de protección de copyright.

3.2 GRUPOS DE ELABORACIÓN DE ESTÁNDARES

Los estándares son la base de la interoperabilidad que garantiza la misión de la infraestructura: la reutilización para minimizar la inversión. Las infraestructuras han de relacionarse con los grupos que proponen estándares.

En la actualidad hay unos 90 estándares que pueden ser relevantes para el PLN y la TA[30]. Hacen referencia a la codificación de la información procesable y a la codificación de la información sobre los datos y recursos en metadatos. Estos estándares, la mayoría consecuencia del esfuerzo por garantizar la interoperabilidad, se han desarrollado en proyectos del PM de la EU con proyectos como EAGLES, PAROLE, SIMPLE y LIRICS, y han sido propuestos como estándares internacionales por la ISO que en su comité WG37 "Terminology and other language and content resources" ha publicado más de 46 estándares. El subgrupo relevante para las tecnologías del lenguaje es el SC4 "Language Resources Management"[31] que ha publicado un total de 19 estándares (actualizaciones incluidas). España participa en este grupo de trabajo mediante el Comité técnico de normalización 191 (AEN/CTN 191), Terminología, presidido por un representante de la Asociación Española de Terminología (AETER).

Las infraestructuras CLARIN y META-SHARE[32] han propuesto estándares en la documentación de recursos y procesadores: los metadatos que han de permitir la búsqueda y que también deberán ser tenidos en consideración en las operaciones para el catálogo de las infraestructuras lingüísticas.

3.3 AGENCIAS DE FINANCIACIÓN

Las agencias de financiación son actores en las infraestructuras en tanto en cuanto implementan las políticas de datos con los actores implicados y determinan la solución de problemas de confidencialidad, protección de datos y uso de licencias de los datos producidos con financiación pública. Deben dictaminar sobre los planes de gestión de datos de los proyectos que financian y las

[30] *Para un listado actualizado de los estándares relevantes ver http://clarin.ids-mannheim.de/standards*
[31] *http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=297592*
[32] *http://www.meta-share.org/p/83/Standards-for-LRs*

infraestructuras lingüísticas serán, por lo menos en el período de construcción, un proyecto financiado con financiación pública. Además, en la implementación del Plan TL, la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información se ha fijado unos objetivos que debe de poder monitorizar y controlar.

3.4 USUARIOS

Por último, las empresas, que son los usuarios objetivo (aunque no los únicos), pymes y start-ups de forma prioritaria porque son los que no pueden autoabastecerse de forma rápida, son participantes en la infraestructura puesto que son los que han de establecer requisitos.

Las empresas del sector de las tecnologías del lenguaje están agrupadas en clústeres más o menos horizontales actualmente: en el País Vasco, Langune[33]; en Cataluña, Clusterlingua[34] y en Madrid, Plataforma del Español[35]. Son asociaciones de empresas que son desarrolladoras o que ya incluyen tecnologías del lenguaje y/o la traducción automática. También existen asociaciones europeas, la más extensa hoy en día es LT-Innovate[36].

La infraestructura debe mantenerse en contacto con los actores cuyo beneficio es el objetivo de la misma: los relacionados con la innovación. Los datos y servicios de la infraestructura han de mantenerse alineados con necesidades de los desarrolladores de tecnología y, en particular, de empresas innovadoras y no necesariamente del ámbito de las tecnologías del lenguaje. Ha de fomentarse la utilización de la infraestructura en la investigación privada, en la elaboración de prototipos y demostradores, y, en último término, en el desarrollo de productos innovadores.

Hemos repasado la lista de potenciales participantes en la gobernanza de las infraestructuras lingüísticas. En la siguiente sección describiremos los antecedentes, tal como se ha presentado en la introducción, para estudiar luego cómo se organiza la gobernanza en cada una de ellas y qué características pueden ser más apropiadas para el caso que nos ocupa.

[33] http://www.langune.com/
[34] http://www.clusterlingua.cat/en/
[35] http://www.plataformadelespañol.com/
[36] http://www.lt-innovate.org/

4 ANTECEDENTES E INICIATIVAS RELACIONADAS

A continuación revisamos el modelo de gobernanza de una serie de entidades que por sus objetivos o por su naturaleza pueden servir como modelos a la gobernanza de las infraestructuras lingüísticas del Plan TL.

A los efectos de la creación de infraestructuras lingüísticas y su gestión, podemos considerar antecedentes a entidades como LDC y ELRA/ELDA que son agencias que recopilan, producen y distribuyen recursos lingüísticos desde finales de los años noventa; más recientemente CLARIN y META-SHARE se han puesto en marcha con vocación de ser infraestructuras europeas. Estas dos últimas se diferencian básicamente en el público objetivo, académico e industrial respectivamente, y en la priorización de algunos objetivos: mientras que META-SHARE se ha preocupado de estandarizar los metadatos y permitir la ubicación y descarga de los datos, CLARIN ha hecho énfasis en aplicar la estandarización de los procesadores y datos para garantizar la interoperabilidad de herramientas. Todas ellas tienen en común el reconocimiento explícito de la importancia de la reutilización de datos y procesadores existentes y de compartir esquemas de evaluación.

También incluimos en esta revisión la iniciativa CREL: *Centre de Referència en Enginyeria Lingüística* que puso en marcha la Generalitat de Cataluña en el período 1995-2000 y cuyo objetivo fue promover la creación de recursos y procesadores para el procesamiento del catalán y asegurar la interoperabilidad entre recursos y procesadores de todos los grupos participantes. El CREL funcionó con un contrato programa que establecía un sistema de gobernanza que puede resultar interesante.

Incluimos también información de la gobernanza de la Organización Internacional para el Desarrollo de Estándares de Terminología Sanitaria (IHTSDO) una asociación de estados dedicada al desarrollo y comercialización de una terminología normalizada para informática clínica y otros productos relacionados.

Por otro lado, también describimos el modelo de gobernanza de tres infraestructuras de otros ámbitos, el BSC-CNS, IRTA y el muy reciente proyecto VISC+. Las tres ofrecen servicios para la innovación y contemplan, en especial VISC+, datos y herramientas, lo que puede ser interesante para establecer paralelismos con las necesidades de las infraestructuras lingüísticas.

4.1 CLARIN

Es el acrónimo de *Common Language Resources and Technological Infrastructure*, una infraestructura de investigación europea que pretende dar acceso fácil y sostenible para los investigadores de las áreas

de humanidades y ciencias sociales a datos lingüísticos y a herramientas que permitan descubrirlos, explorarlos, explotarlos, anotarlos y analizarlos, con independencia de dónde estén ubicados.

CLARIN está construyendo una federación de repositorios de datos, centros de servicios y centros de expertos con un único acceso para todos los miembros de la comunidad. Las herramientas y los datos convergen gracias a esquemas de interoperabilidad que permiten la combinación de los mismos.

CLARIN ofrece servicios[37] relacionados con:

1. un catálogo para la búsqueda y descubrimiento de conjuntos de datos, cómo acceder a ellos y guías con definición de estándares y planes de gestión de datos que cubren todo su ciclo de vida para las tareas de creación de nuevos conjuntos de datos;

2. servicios web para minería y análisis de datos lingüísticos;

3. descarga de procesadores de anotación y explotación; y

4. archivo y almacenamiento de datos y políticas y herramientas para compartir datos. Es particularmente importante su contribución a la estandarización técnica de procesos (ver la sección 5) y a la clasificación de licencias[38] según los usos que favorezcan la reutilización. CLARIN se ha propuesto también (en LT-Observatory) para mediar en el acceso comercial a datos lingüísticos[39].

[37] https://www.clarin.eu/content/services
[38] https://www.clarin.eu/content/license-categories
[39] http://www.lrec-conf.org/proceedings/lrec2016/pdf/1150_Paper.pdf

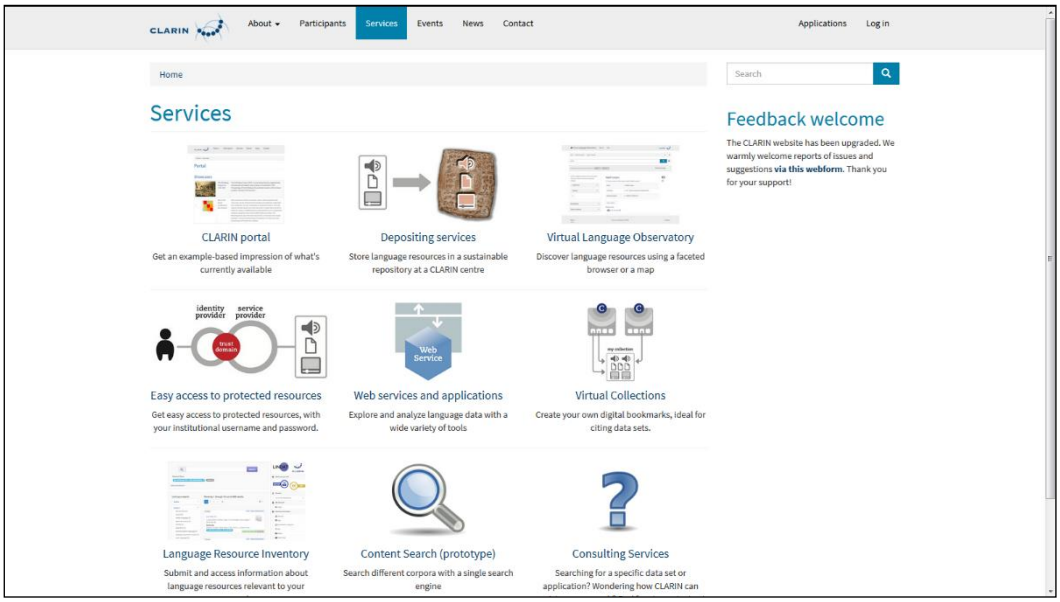


Figura 2: Captura de pantalla de la página que da acceso a los servicios ofrecidos por CLARIN (acceso 22-11-2016)

Figura 2: Captura de pantalla de la página que da acceso a los servicios ofrecidos por CLARIN (acceso 22-11-2016)

La infraestructura está dotada de un órgano de gobernanza y coordinación a nivel europeo, la CLARIN ERIC, un *European Research Infrastructure Consortium*[40] cuyos miembros son países u organizaciones intergubernamentales.

La *Asamblea General* es el órgano superior de decisión de CLARIN-ERIC, y está integrada por representantes ministeriales de los países miembros. La Asamblea General está asistida por el *Consejo Científico Internacional Asesor*.

La gestión operativa y la implementación de las estrategias y normativa establecidas por la Asamblea General está delegada en la *Junta de Directores*, que cuenta con el apoyo de la *Oficina CLARIN*, cuya entidad jurídica es la Facultad de Humanidades de la Universidad de Utrecht (Holanda). La Junta de Directores consta de cinco miembros: Director/a Ejecutiva, Subdirector/a Ejecutiva, el presidente del Foro de Coordinadores Nacionales, el Director de "Participación de usuarios" y el Director Técnico de la infraestructura.

[40] *"Being considered as an international organisation within the meaning of the directive on public procurement (Directive 2004/18/EC and Directive 2014/24/EC)"*

El *Foro de Coordinadores Nacionales*, que consta de los coordinadores de los consorcios nacionales, 19 en el momento de escribir este informe, tiene como responsabilidad garantizar la coordinación de la implementación de las estrategias propuestas por la Asamblea General.

El *Comité Permanente de Centros Técnicos CLARIN*, compuesto a su vez por los directores de los centros nacionales principales, es el órgano donde se gestiona y garantiza la integración y la interoperabilidad de los componentes de la infraestructura de y para todos los centros y de todos los países miembros.

El *Director de Participación de usuarios* tiene como principal responsabilidad el fomento de la utilización de la infraestructura, así como la coordinación para recabar requerimientos de los usuarios.

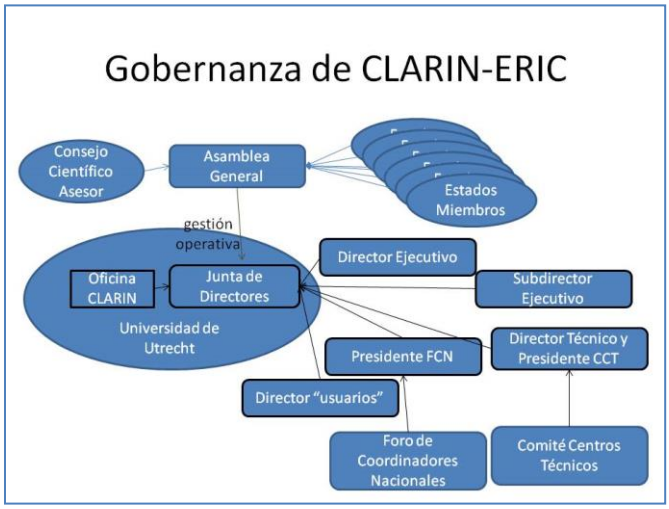


Figura 3: Diagrama de la relación entre los órganos de planificación y operaciones del ERIC CLARIN

4.1.1 CLARIN EN LOS ESTADOS MIEMBROS

La gobernanza de las federaciones que constituyen los miembros es de diversa índole. Por ejemplo, CLARIN-D, financiado por el Ministerio de Educación e Investigación alemán, que se ha organizado como *una red de centros*, en la que cada uno ofrece un área de especialización, relacionándose a nivel nacional con asociaciones de usuarios y centros de computación. La red tiene un Comité Asesor internacional formado por siete expertos de reconocido prestigio.

En el caso de Holanda, CLARIN-NL ha podido reutilizar la estructura del programa STEVIN (Acrónimo para Recursos y Tecnologías de la Lengua y del Habla Esenciales)[41]. Este fue un programa que se cerró

[41] http://www.efnil.org/documents/conference-publications/thessaloniki-2010/language-languages-and-new-technologies/12-Cucchiarini-Bosch.pdf

en 2011 y cuyo objetivo era estimular la producción de recursos y tecnología para el holandés. CLARIN-NL se creó en 2012 y en 2014 se unió a otra infraestructura para humanidades DARIA creando CLARIAH-NL[42]. Está formado por 24 universidades y centros de investigación, de los cuales seis fueron los fundadores. Tiene también relaciones con empresas a los que presta servicios de apoyo a la innovación y el desarrollo. Este consorcio se gobierna mediante un Consejo formado por nueve miembros. De estos cuatro constituyen el Consejo Ejecutivo, responsable de la gestión de operaciones, tres de los miembros son los Project Managers de las tres áreas en las que se centra la infraestructura. Un miembro más representa al sector informático y otro al colectivo DARIAH.

Siendo CLARIN una infraestructura europea, pretende implicar a todos los estados miembros de la Unión. En el caso de España, el Ministerio de Ciencia e Innovación publicó en 2010 la estrategia española con respecto a su participación en las infraestructuras europeas. En este informe catalogó CLARIN de "interés medio" para España y rechazó participar, a pesar de que había una red de centros ya organizada gracias a la financiación europea, del propio ministerio y de los gobiernos autonómicos, en particular de la Generalitat de Cataluña. A pesar de no pertenecer a la infraestructura, el nodo inicial de la infraestructura en España sigue, en forma de Centro de conocimiento-Español (Spanish CLARIN-CentreK[43]), participando y relacionándose con la infraestructura europea.

4.2 META-SHARE

META-SHARE es una red de repositorios, abierta y segura, establecida para compartir e intercambiar datos lingüísticos, herramientas y servicios relacionados. Su objetivo es asegurar la sostenibilidad de la difusión y el intercambio de recursos lingüísticos para las tecnologías del lenguaje y otros ámbitos donde el lenguaje juega un papel crucial. META-SHARE se ha implementado como un catálogo (basado en un rico sistema de metadatos) que permite subir y descargarse recursos y herramientas y acceder a APIs que cubren:

- Datos lingüísticos, tales como corpus escritos y orales
- Datos relacionados con los datos lingüísticos tales como recursos multimedia
- Procesadores y anotadores lingüísticos
- Servicios para utilizar directamente los procesadores y anotadores
- Herramientas de evaluación, métricas y protocolos y los servicios dirigidos a la evaluación

[42] http://www.clariah.nl
[43] https://centres.clarin.eu/centre/32

- Flujos de trabajo (pipelines o workflows) que combinan diferentes servicios interoperables

META-SHARE es una infraestructura cuyos objetivos son los siguientes y entre sus contribuciones destaca una revisión exhaustiva de los esquemas de licencias relevantes para el ámbito con una propuesta específica y motivada de licencias[44].

1. Dar acceso a recursos lingüísticos de calidad y sus metadatos
2. Preservar y mantener los recursos y sus metadatos
3. Promocionar la utilización de estándares para el desarrollo de recursos lingüísticos con el objetivo de garantizar la máxima interoperabilidad
4. Proporcionar una serie de servicios para los miembros y usuarios de META-SHARE
5. Permitir la fácil indexación de RL de terceros a través de la red
6. Permitir la descarga y adquisición legal de los RL de calidad que ofrece.

META-SHARE está vinculada a la red de excelencia META-NET[45]. La participación en META-SHARE está regulada por el *META-SHARE Charter*[46] y por un *Memorandum of Understanding*[47]. En este documento se relacionan los diferentes tipos de socios, a saber:

- META-SHARE Network Nodes son miembros de la red que disponen de un repositorio META-SHARE y lo mantienen y actúan como proveedores de servicios del repositorio.
- META-SHARE Managing Nodes son miembros de la red que son responsables de los servicios centrales para todos los miembros y actúan como soporte de los usuarios proveedores de servicios.
- META-SHARE Depositors son miembros de la red que depositan sus recursos en un repositorio de un META-SHARE Network Node.

El *Comité Ejecutivo* (Executive Board) de la red META-NET es el órgano asesor en cuestiones fundamentales de la dirección de META-SHARE.

[44] http://www.meta-share.org/p/82/Legal-issues
[45] www.meta-net.eu
[46] http://www.meta-share.org/assets/pdf/METASHARE_Charter.pdf
[47] http://www.meta-share.org/assets/pdf/META-SHARE_MoU_v2.1.pdf

La coordinación de las operaciones de META-SHARE es asumida por el *Comité de Coordinación* META-SHARE, que consta de 5 miembros, designados y nombrados por mayoría simple entre los administradores de repositorios y nodos gestores, que son nombrados por cada uno de los nodos de la red. Los administradores de repositorio y representantes de los nodos gestores son responsables de mantener el inventario y asegurar que los LR depositados están en conformidad con las normas de protección de datos y derechos de propiedad intelectual, así como con las políticas vigentes de la organización respectiva. El Comité de Coordinación tiene un mandato de un año.
El Comité de Coordinación META-SHARE decide por mayoría simple en las cuestiones operativas y técnicas, incluyendo:

- Aceptación de nuevos miembros META-SHARE para actuar como proveedores de servicios de repositorio META-SHARE y de nuevos nodos gestores META-SHARE para actuar como proveedores de servicios básicos y de apoyo al usuario;
- Delegación de la prestación de determinados servicios a un miembro META-SHARE;
- Modificaciones en la configuración técnica y esquemas de licencias de la Red META-SHARE;
- Cuestiones derivadas del incumplimiento del nivel de servicios acordado para los servicios básicos META-SHARE y los repositorios de META-SHARE.

El Comité de Coordinación en su conjunto, o cualquiera de sus miembros, puede ser renombrado sin restricciones. El Comité de Coordinación nombra entre los representantes de los nodos gestores a su Presidente, por un plazo de un año y que puede ser renovado hasta un máximo de tres veces.

4.3 ELRA/ELDA

La *European Association for Language Resources* se registró en Luxemburgo en febrero de 1995. En los primeros años ELRA recibió apoyo de la Comisión Europea a través de la financiación de proyectos, pero ha sido autosuficiente desde 1998. Paralelamente y también en 1995 se creó la *Evaluations and Language resources Distribution Agency* (ELDA) como la unidad operativa de ELRA. ELDA está encargada del desarrollo y la ejecución de la misión y tareas de ELRA definidas por la Junta de la asociación. ELDA está constituida como una empresa con el objetivo de dotarla de los medios para llevar a cabo todas las tareas relacionadas con la comercialización de recursos y la prestación de servicios relacionados. El director general de la agencia, Khalid Choukri, es también el Secretario General de ELRA.

La misión de ELRA es ser una organización central, sin ánimo de lucro, que vele por la recopilación, distribución y validación de recursos de voz, texto, terminologías y procesadores asociados. ELRA se ha ocupado, desde sus inicios, en abordar mediante comités específicos de trabajo los problemas de diversa naturaleza, técnicos y logísticos, cuestiones comerciales (precios, cuotas, regalías), problemas legales (licencias, Derechos de Propiedad Intelectual) relacionados con los recursos del lenguaje y su evaluación así como sobre la difusión de información relacionada. La Asociación es la organizadora de la Conferencia Internacional Language Resources and Evaluation, LREC, que ha celebrado ya su X edición, en formato bianual, y que está considerado uno de los congresos más influyentes del ámbito (Google Scholar lo situa en el puesto 5 del ranking de publicaciones del área Lingüística Computacional, y las actas de diferentes ediciones están indexadas por Thomson Reuters[48]). La revista de la asociación es Language Resources and Evaluation (LRE) publicada por Springer e indexada por Thomson Reuters en el JCR.

ELRA (en colaboración con ELDA) mantiene diferentes catálogos de recursos:

ELRA Catalogue: http://catalog.elra.info/

META-SHARE: http://metashare.elda.org/

LRE-Map: http://www.elra.info/en/catalogues/lre-map/

R&D Catalogue: http://www.elra.info/en/catalogues/r-and-d-catalogue/

Universal Catalogue: http://www.elra.info/en/catalogues/universal-catalogue/

ELRA también ofrece los siguientes servicios:

1. Identificación, colección y distribución de recursos lingüísticos
2. Producción por encargo de recursos lingüísticos
3. Evaluación y validación de recursos lingüísticos
4. Estandarización de recursos lingüísticos
5. Consultoría para productores de recursos sobre cuestiones legales y de derechos de propiedad intelectual en torno a recursos lingüísticos
6. Evaluación de procesadores lingüísticos

[48] *http://lrec2016.lrec-conf.org/en/about/conference-proceedings/*

7. Promoción de los recursos y procesadores lingüísticos
8. Consultoría y estudios disponibilidad de recursos y tecnología para una lengua y los requerimientos necesarios
9. Consultoría y asesoría en planificación de desarrollos lingüísticos.

Recientemente ELRA ha liderado la implantación del *International Standard Language Resource Number* (ISLRN) un esquema de identificador único y universal para los recursos lingüísticos.

ELRA está gobernada por la *Asamblea General* formada por un representante de cada miembro de la asociación. Los miembros de la asociación son entidades legales de carácter público y privado y solamente tienen voto los miembros que pertenecen a estados europeos. Los no-europeos pueden asociarse como *Subscribers*, con todos los derechos, pero sin voto en la Asamblea. La Asamblea General tiene un Presidente y está asistida por un Secretario.

La gobernanza de la asociación se delega en un Consejo formado por nueve miembros elegidos por votación por la Asamblea para un período de dos años, que puede renovarse hasta un máximo de seis años. El Consejo elige entre sus miembros un presidente, un vicepresidente, un secretario y un tesorero. El presidente actúa como el representante de la asociación en todos los asuntos y puede delegar en el vicepresidente o en otro miembro del Consejo.

El Consejo debe elegir un Comité de Ética formado por, por lo menos, cuatro miembros del Consejo incluido el Presidente. Su papel es asesorar al Presidente sobre todas las cuestiones relacionadas con potenciales conflictos de interés entre los miembros del Consejo y la Asociación o ELDA.

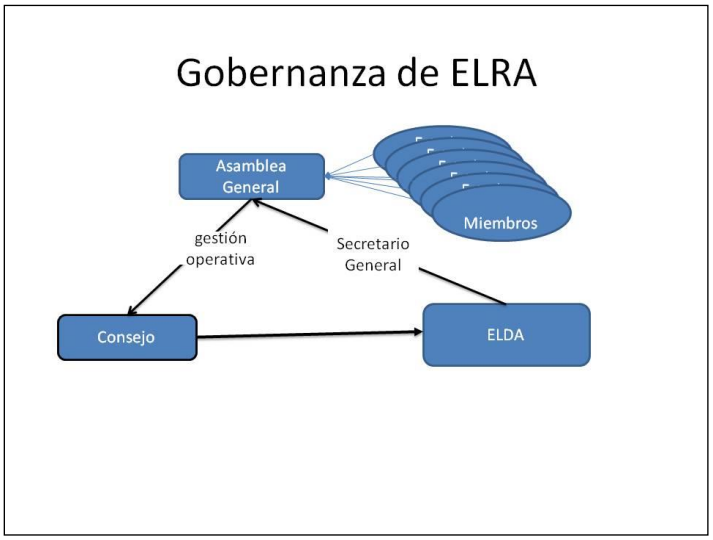


Figura 4: Diagrama de la relación entre los órganos de planificación y operaciones de la European Language Resources Association

4.4 LDC

El *Linguistic Data Consortium* (LDC) es un consorcio de universidades, bibliotecas, empresas y laboratorios de investigación de los EUA. Se formó en 1992 primero para mitigar la escasez de datos lingüísticos críticos para la investigación y el desarrollo de las tecnologías del lenguaje. La *Advanced Research Projects Agency* (ARPA) proporcionó inicialmente la financiación y la National Science Fundation (NSF) a través de la Dirección de Información y Sistemas Inteligentes aportó también financiación en forma de proyectos. Según su declaración de intenciones, el principio fundacional del LDC es que el acceso a datos es el motor de la innovación.

Inicialmente, el papel principal del LDC era actuar de repositorio de los recursos lingüísticos y encargarse de su distribución. LDC ha crecido hasta convertirse en una organización autofinanciada que ofrece como servicios básicos crear y distribuir una amplia gama de datos lingüísticos. LDC también patrocina proyectos de investigación y evaluaciones de tecnología basadas en el lenguaje, proporcionando recursos y asesoría.

La organización LDC está ubicada en la Universidad de Pennsylvania como un centro dentro de la Escuela de Artes y Ciencias de la Universidad, y su director es Mark Liberman, profesor del departamento de Lingüística y del departamento de Ciencias de la computación y la información. El Director ejecutivo (actualmente Christopher Cieri) tiene encomendadas las tareas de planificación, operaciones, gestión de proyectos, relaciones externas y la gestión económica. Cuenta en la actualidad con más de 40 trabajadores.

4.5 CREL

El Centre *de Referència de Recerca i Desenvolupament en Enginyeria Lingüística de la Generalitat de Catalunya* se constituyó en 1996 y se extinguió en 2010. Estaba integrado por nueve grupos de diferentes entidades que comprendían grupos de investigación de las diferentes universidades catalanas y el Institut d'Estudis Catalans (IEC), que actuó como entidad gestora del centro.

Los objetivos del centro eran: llevar a cabo investigación y desarrollo en el área del procesamiento del catalán basándose en la complementariedad de los grupos que lo constituían; potenciar las infraestructuras de I+D en ingeniería lingüística evitando la duplicación de esfuerzos y crear servicios de apoyo para grupos de investigación públicos y privados; captar financiación exterior, nacional e internacional, y participar en redes con objetivos similares; intervenir activamente en tareas de formación de interés para empresas y organismos públicos de investigación; desarrollar la transferencia de servicios y tecnologías; difundir los resultados de la investigación y asesorar al gobierno de la Generalitat y a los organismos e instituciones que dependen de ésta en los temas relacionados con su ámbito de actuación y especialidades.

El Centro, sin perjuicio de la representación que tenían las entidades titulares de cada uno de los grupos participantes, se dotó de los siguientes órganos de gobernanza:

- *Entidad Gestora*: ostentaba la representación legal del centro delante de terceros, acogía la sede del centro y ejercía la actividad contractual y las facultades de administración de los bienes y derechos del centro.
- *Consejo de dirección*: integrado por el presidente, que era el director del centro, tres representantes del Consejo científico, dos representantes de la Generalitat, y el secretario que era el representante de la entidad gestora. Este Consejo de dirección era responsable de proponer programas anuales y hacer el seguimiento y control de los mismos, proponía el presupuesto anual del CREL a la Generalitat, controlaba los recursos asignados y proponía actividades a los grupos integrantes del centro.
- *Consejo científico*: integrado por el director, un representante de cada uno de los equipos integrantes, y un representante de cada centro asociado, tenía como funciones: coordinar el desarrollo y las actividades del CREL en función de los objetivos del contrato programa; proponía al Consejo de dirección las líneas de investigación y desarrollo tecnológico; velaba por el uso óptimo de los recursos humanos y los equipos del CREL; proponía acciones de

formación, de transferencia de tecnología y de difusión; proponía al Consejo de dirección la aceptación de centros asociados y colaboradores.

- *Director*, que era nombrado por la Generalitat a propuesta del Consejo de dirección, tenía las funciones de: coordinar el funcionamiento, informar al Consejo de dirección del funcionamiento del centro; ejecutar los acuerdos del Consejo de dirección y del Consejo científico, ejercer las tareas de administración que el Consejo de dirección delegaba; representar institucionalmente al centro.

El control científico y técnico de las actuaciones se llevaba a cabo mediante la realización de auditorías científicas.

4.6 IHTSDO

La Organización Internacional para el Desarrollo de Estándares de Terminología Sanitaria (International Health Terminology Standards Development Organization, IHTSDO) es la organización internacional sin ánimo de lucro propietaria de los derechos de propiedad intelectual de SNOMED-CT, un recurso terminológico para medicina clínica que comercializa como producto junto con su modelo conceptual y las licencias de programas de software relacionados con la misma. La IHTSDO produce y mantiene este recurso terminológico en inglés y español, pero ofrece ayudas para su traducción a otras lenguas, o a variantes de las mismas, por ejemplo, existe una extensión a la versión en español europeo. Además, desarrolla y licencia diferentes programas de gestión de la terminología. La IHTSDO solamente cobra las licencias a las entidades que no pertenecen a un estado miembro.

Se creó en 2007 formada por nueve países que son los miembros fundadores de la organización. Actualmente tiene más de 20 miembros que pueden ser países, naciones, estados o zonas geográficas que tengan derecho a voto en la Organización de las Naciones Unidas, no pudiendo existir más de un miembro para o con respecto a un único territorio. España forma parte de la IHTSDO desde 2009.

El máximo órgano de gestión de la IHTSDO es la Asamblea General, en la que cada miembro tiene un voto. El consejo de administración ejerce la gestión y dirección de la IHTSDO y es elegido por la Asamblea General, no obstante, los directores que forman parte del consejo de administración no representan a sus países. Además, cuenta con un Equipo de gestión, liderado por un director ejecutivo, y unos diez diferentes jefes de operaciones en diferentes áreas.

Además, dispone de un Foro de miembros, que es un órgano consultivo cuyo rol es facilitar la colaboración y cooperación entre los miembros y es el canal de comunicación entre la asociación y sus

miembros. Está presidido por un representante del consejo de administración y un representante de los miembros de la asociación.

El Foro de enlace con proveedores asesora al consejo de administración y hace de canal para que las opiniones y sugerencias de los proveedores sobre el desarrollo, versiones e implementación de SNOMED CT. Además, se organiza en diferentes grupos de trabajo de carácter consultivo para temas concretos de la gestión de la terminología.

4.7 BSC-CNS

El Consorcio Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-BCN) es una infraestructura de supercomputación para la investigación básica e investigación aplicada. Tiene relaciones de colaboración con empresas como IBM, intel, Repsol y Nvidia. El centro BSC-CNS es un Centro de Supercomputación Nacional que coordina la Red Española de Supercomputación (RES). Se constituyó en 2005 como una entidad de derecho público con participación de la Administración General del Estado con personalidad jurídica propia. "Su objeto o finalidad es la de gestionar y promover la colaboración científica, técnica, económica y administrativa de las Instituciones que lo integran, para la creación, construcción, equipamiento y explotación del BSC-CNS, como centro de servicios de supercomputación para uso multidisciplinar, abierto a la comunidad nacional de científicos y técnicos, de entidades públicas y privadas, orientado para fomentar la colaboración internacional, conectado a través de las redes de comunicaciones a otros centros e instituciones de su ámbito, con un Proyecto Científico y Tecnológico inicial que contiene sus objetivos a medio plazo, los medios necesarios para su ejecución y su propia estructura orgánica y funcional." [49]

El esquema de gobernanza del Barcelona Supercomputing Centre es el que se representa en la Figura 5.

[49] *BOE-A-2016-2170*

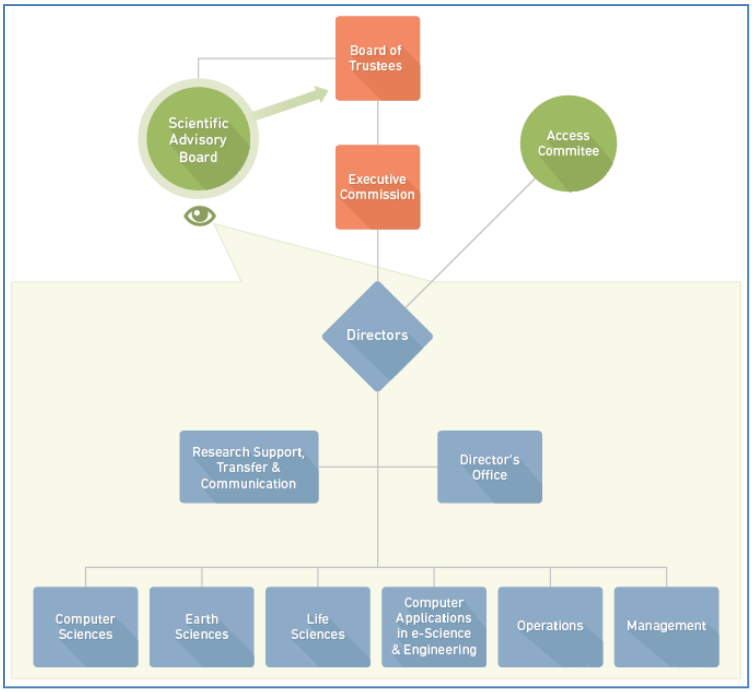


Figura 5: Diagrama de la relación entre los órganos de planificación y operaciones del BSC-CNS que figura en su página web

Los órganos de gobernanza del BSC-CNS son: *el Consejo Rector* y la *Comisión Ejecutiva* que hace el seguimiento y ejecuta todas las actividades del centro.

El Consejo Rector, que está asesorado por un consejo científico, está formado por representantes del Ministerio de Economía y Comercio, de la Generalitat de Catalunya y de la Universitat Politècnica de Catalunya y del BSC-CNS: el director, el director asociado, el Gerente, y dos Secretarios (Abogados del Estado).

El Consejo Rector nombra al Director del centro y al comité encargado de decidir sobre el uso científico del supercomputador MareNostrum: es el Comité de Acceso formado por expertos españoles externos al centro, la mitad de ellos nombrados por la ANEP.

Los principales órganos de gestión del Consorcio BSC-CNS son la Junta de Directores y la Comisión Ejecutiva compuestos por miembros de las instituciones que participan en el consorcio: el Ministerio de Educación y Ciencia, Departament d'Educació i Universitats de la Generalitat de Catalunya y la Universitat Politècnica de Catalunya.

La Junta de Directores se ocupa de la gestión de operaciones del centro y el seguimiento de las funciones técnicas y científicas y está formada por el director, el director asociado, el director de

operaciones, el gerente y los directores de cada área del centro: Ciencias de la computación, Ciencias de la vida, Ciencias de la tierra, Aplicaciones para e-ciencia e ingeniería.

La Comisión Ejecutiva está formada por su presidente, que es un representante del Ministerio, su vicepresidente, que es un representante de la Generalitat, junto con un vocal del Ministerio, un vocal de la Generalitat, dos vocales de la UPC y el Director, el Director Asociado, el Gerente y como secretarios dos abogados del Estado.

El Comité de Acceso al MareNostrum y la Red RES está constituido por cuatro miembros: un gestor externo al BSC-CNS, un experto en supercomputación nombrado por la ANEP, un experto en supercomputación externo al BSC-CNS y un miembro del BSC-CNS. A su vez, el Comité de Acceso está asesorado por un Panel de Expertos, científicos españoles de reconocido prestigio externos al BSC. El Panel está dividido en cuatro grupos siguiendo la clasificación de la FECYT: astronomía, espacio y ciencias de la tierra; biomedicina y ciencias de la salud; física e ingeniería; química y materiales y tecnología. El panel de expertos puede requerir la asistencia de la ANEP para llevar a cabo evaluaciones por pares de los proyectos de utilización que recibe.

4.8 IRTA

El Instituto de Investigación y Tecnología Agroalimentarias (IRTA) es una entidad de derecho público con personalidad jurídica propia, y adscrito al departamento competente en materia de agricultura y alimentación de la Generalitat de Cataluña. La finalidad del Instituto de Investigación y Tecnología Agroalimentarias, de acuerdo con las directrices de las políticas agroalimentaria y de investigación, desarrollo y transferencia (I+D+T) del Gobierno y del departamento competente en materia de agricultura y alimentación, es contribuir a la modernización, mejora e impulso de la competitividad; al desarrollo sostenible de los sectores agrario, alimentario, agroforestal, acuícola y pesquero, así como de los directa o indirectamente relacionados con el abastecimiento de alimentos sanos y de calidad a los consumidores finales; a la seguridad alimentaria y a la transformación de los alimentos, y, en general, a la mejora del bienestar y la salud de la población.

Está formado por 10 centros consorciados. El conjunto de centros con finalidades de investigación, desarrollo y transferencia en que participa el Instituto constituye el sistema cooperativo de investigación y desarrollo agroalimentarios. Dichos centros se consideran centros concertados con el Instituto, con el que mantienen los vínculos orgánicos, operativos y financieros que los respectivos órganos de gobierno puedan establecer. Como centros concertados, reciben apoyo del Instituto y

pueden ser incluidos en las previsiones y acuerdos del contrato-programa con el Gobierno. La participación del Instituto en los diferentes órganos de gobierno de dichos centros concertados se lleva a cabo de acuerdo con las decisiones del Consejo de Administración del Instituto.

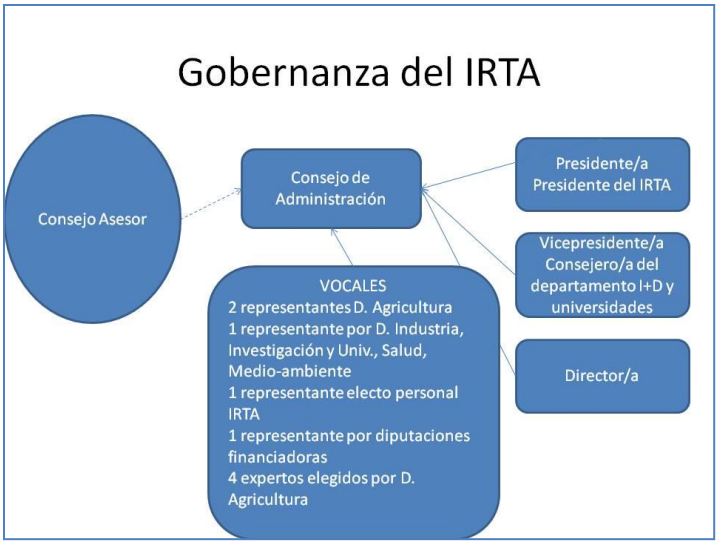


Figura 6: Diagrama de la relación entre los órganos de planificación y operaciones del Institut de Recerca en Tecnologia dels Aliments

4.9 PROYECTO VISC+

El proyecto VISC+ tiene como objetivo ofrecer la información de salud que se genera en Cataluña, de forma totalmente anonimizada y segura, para impulsar y facilitar la investigación, la innovación y la evaluación en los ámbitos de medicina y ciencias de la salud. El proyecto está orientado a los investigadores de centros públicos de investigación y los agentes del sistema sanitario (SISCAT), que verán facilitada su tarea en lo que se refiere a la recogida, tratamiento y análisis de datos de salud.

El proyecto dispone de una gran base de datos de salud. Estos datos serán procesados, especialmente para la anonimización y desidentificación de datos personales y los suministrará para investigaciones concretas. El proyecto dará servicios como: elaboración y suministro de conjuntos de datos, análisis estadístico de conjuntos de datos y preparación de informes.

La operativa y la gestión de VISC+ está delegada en la Agencia de Calidad y Evaluación Sanitaria de Cataluña (AQUAS), que es una entidad de derecho público adscrita al Departamento de Salud de la Generalitat de Catalunya, que actúa al servicio de las políticas públicas del departamento, y que está sometida al ordenamiento jurídico privado. La AQUAS tiene la misión de generar conocimiento relevante para contribuir a mejorar la calidad, la seguridad y la sostenibilidad del sistema de salud

catalán. La AQUAS, mediante un contrato de servicios resultante de un proceso de diálogo competitivo, delegará en un colaborador externo el desarrollo de algunos de los componentes necesarios para la operativa (una plataforma de análisis de datos, estadísticas, etc.).

El proyecto contará con los siguientes órganos de gobernanza (el proyecto está parado después de una fuerte controversia sobre el uso de datos[50]). Un *Consejo Científico Asesor*, formado por expertos internacionales en materia de investigación biomédica y docencia en estadística y análisis de datos, tiene como misión garantizar la calidad de los servicios y productos que se realicen y asesorar sobre las demandas de uso de los investigadores e investigadores potenciales. El proyecto se adscribirá a un Comité Ético de Investigación Científica de los ya existentes. La misión será velar por la correcta aplicación de los criterios científico-éticos en la gestión de las peticiones de datos y servicios y aportar la aprobación que la normativa vigente requiere.
La AQUAS se habrá de dotar de un comité de gobierno que se encargue del seguimiento del proyecto y que tendrá responsabilidad sobre los siguientes puntos: liderazgo institucional; estrategia global del proyecto y toma de decisiones clave; aprobación del código ético, políticas de transparencia y de seguridad y del proceso de anonimización/desidentificación llevado a cabo por AQUAS.

5 ESTÁNDARES EN LAS TECNOLOGÍAS DEL LENGUAJE

La infraestructura lingüística tiene que servir al objetivo de la reutilización: la reutilización de datos para entrenar diferentes procesadores, y reutilización de procesadores para producir nuevos datos. Es necesario entonces facilitar, en la medida de lo posible, la reutilización mediante la selección de especificaciones técnicas y su aplicación de forma que se conviertan en estándares. La utilización de estándares facilita también la evaluación y ésta es un potente factor de normalización. Por último, la utilización de estándares facilita el desarrollo de aplicaciones puesto que no es necesario adaptar las herramientas para cubrir diferentes formatos en el caso de usar diferentes procesadores, por ejemplo, debido a diferentes lenguas.

A modo de ejemplo, en las siguientes figuras copiamos el código generado para el análisis morfosintáctico de la misma oración en inglés "the girls couldn't attend the party " por tres cadenas de procesadores diferentes: FreeLing, IxaPipes y Stanford CoreNLP.

[50] http://aquas.gencat.cat/ca/actualitat/proce_participatiu_deliberatiu_analitica_dades_salut/ y http://ccaa.elpais.com/ccaa/2016/06/02/catalunya/1464863725_879219.html

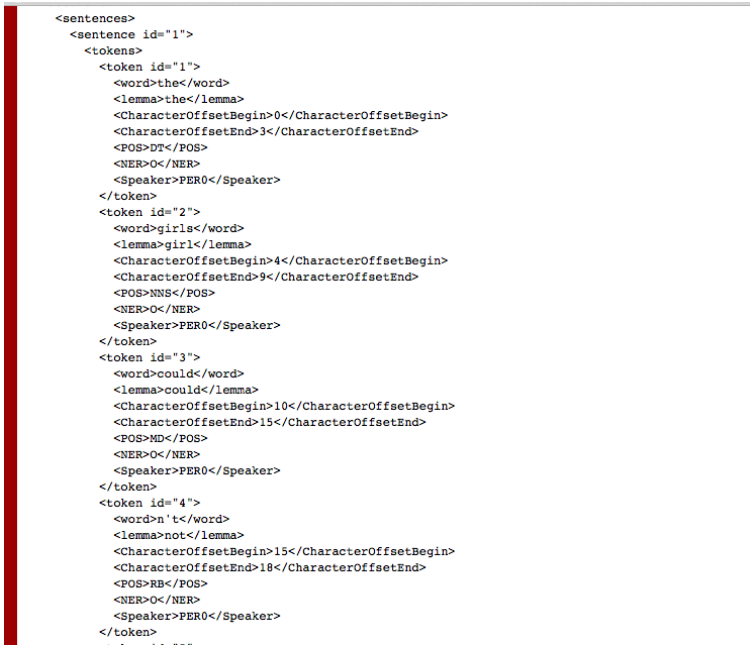
▼ XML output

```
<document>
  <wordcount>8</wordcount>
  <cputime>0.003814</cputime>
  <paragraph>
    <sentence id="1">
      <token ctag="DT" form="the" id="t1.1" lemma="the" pos="determiner" tag="DT">
        <morpho>
          <analysis ctag="DT" lemma="the" pos="determiner" selected="1" tag="DT"/>
        </morpho>
      </token>
      <token ctag="NNS" form="girls" id="t1.2" lemma="girl" num="plural" pos="noun" tag="NNS" type="common">
        <morpho>
          <analysis ctag="NNS" lemma="girl" num="plural" pos="noun" selected="1" tag="NNS" type="common"/>
        </morpho>
      </token>
      <token ctag="MD" form="could" id="t1.3" lemma="can" pos="verb" tag="MD" type="modal">
        <morpho>
          <analysis ctag="MD" lemma="can" pos="verb" selected="1" tag="MD" type="modal"/>
        </morpho>
      </token>
      <token ctag="RB" form="not" id="t1.4" lemma="not" pos="adverb" tag="RB" type="general">
        <morpho>
          <analysis ctag="RB" lemma="not" pos="adverb" selected="1" tag="RB" type="general"/>
        </morpho>
      </token>
      <token ctag="VB" form="attend" id="t1.5" lemma="attend" pos="verb" tag="VB" vform="infinitive">
        <morpho>
          <analysis ctag="VB" lemma="attend" pos="verb" selected="1" tag="VB" vform="infinitive"/>
          <analysis ctag="VBP" lemma="attend" pos="verb" tag="VBP" vform="personal"/>
        </morpho>
      </token>
      <token ctag="DT" form="the" id="t1.6" lemma="the" pos="determiner" tag="DT">
```

Figura 7: Resultado del análisis morfosintáctico de FreeLing en formato XML

```
  <raw>the girls couldn't attend the party.</raw>
  <text>
    <wf id="w1" sent="1" para="1" offset="0" length="3">the</wf>
    <wf id="w2" sent="1" para="1" offset="4" length="5">girls</wf>
    <wf id="w3" sent="1" para="1" offset="10" length="5">could</wf>
    <wf id="w4" sent="1" para="1" offset="15" length="3">n't</wf>
    <wf id="w5" sent="1" para="1" offset="19" length="6">attend</wf>
    <wf id="w6" sent="1" para="1" offset="26" length="3">the</wf>
    <wf id="w7" sent="1" para="1" offset="30" length="5">party</wf>
    <wf id="w8" sent="1" para="1" offset="35" length="1">.</wf>
  </text>
  <terms>
    <!--the-->
    <term id="t1" type="close" lemma="the" pos="D" morphofeat="DT">
      <span>
        <target id="w1"/>
      </span>
    </term>
    <!--girls-->
    <term id="t2" type="open" lemma="girl" pos="N" morphofeat="NNS">
      <span>
        <target id="w2"/>
      </span>
    </term>
    <!--could-->
    <term id="t3" type="close" lemma="can" pos="O" morphofeat="MD">
      <span>
        <target id="w3"/>
      </span>
    </term>
    <!--n't-->
    <term id="t4" type="open" lemma="not" pos="A" morphofeat="RB">
      <span>
        <target id="w4"/>
```

Figura 8: Resultado del análisis morfosintáctico de IxaPipes en formato XML



```
<sentences>
  <sentence id="1">
    <tokens>
      <token id="1">
        <word>the</word>
        <lemma>the</lemma>
        <CharacterOffsetBegin>0</CharacterOffsetBegin>
        <CharacterOffsetEnd>3</CharacterOffsetEnd>
        <POS>DT</POS>
        <NER>O</NER>
        <Speaker>PER0</Speaker>
      </token>
      <token id="2">
        <word>girls</word>
        <lemma>girl</lemma>
        <CharacterOffsetBegin>4</CharacterOffsetBegin>
        <CharacterOffsetEnd>9</CharacterOffsetEnd>
        <POS>NNS</POS>
        <NER>O</NER>
        <Speaker>PER0</Speaker>
      </token>
      <token id="3">
        <word>could</word>
        <lemma>could</lemma>
        <CharacterOffsetBegin>10</CharacterOffsetBegin>
        <CharacterOffsetEnd>15</CharacterOffsetEnd>
        <POS>MD</POS>
        <NER>O</NER>
        <Speaker>PER0</Speaker>
      </token>
      <token id="4">
        <word>n't</word>
        <lemma>not</lemma>
        <CharacterOffsetBegin>15</CharacterOffsetBegin>
        <CharacterOffsetEnd>18</CharacterOffsetEnd>
        <POS>RB</POS>
        <NER>O</NER>
        <Speaker>PER0</Speaker>
      </token>
```

Figura 9: Resultado del análisis morfosintáctico de Stanford CoreNLP en formato XML

La comunidad de PLN ha mantenido una intensa actividad en la definición de especificaciones para la información lingüística que enriquece los diferentes tipos de recursos (corpus, léxico y terminología, monolingüe y bilingüe) y que es utilizada por los procesadores. Los procesadores deben adecuar sus mecanismos para input/output para aceptar y entregar, al menos, datos en el formato estándar definido.

Las especificaciones y estándares que empezaron con los trabajos del grupo EAGLES (*European Advisory Group on Language Engineering Standards*) han continuado en el grupo WG37 de la ISO, SC4 SC4 "Language Resources Management"[51] que ha publicado un total de 19 estándares (actualizaciones incluidas) con resultados, en cuanto a adopción se refiere, desiguales. Es importante señalar que hay una actividad internacional importante para marcar estándares para cada lengua. No obstante, se está imponiendo –con la fuerza de que Stanford Core NLP y Google[52] lo promueven y utilizan desde 2012-- la visión de que aun perdiendo información lingüística (o reduciendo calidad de análisis) es mejor definir un estándar "universal"[53] que facilite que determinadas aplicaciones puedan consumir datos de cualquier lengua. Esta propuesta responde al problema que causa para el desarrollo de aplicaciones

[51] http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=297592

[52] https://cloud.google.com/natural-language/

[53] http://universaldependencies.org/#language-u y https://github.com/slavpetrov/universal-pos-tags con extensiones para cada lengua.

multilingües (es decir que desde el diseño puedan aplicarse a textos en diferentes lenguas sin necesidad de revisar el código dependiendo de la lengua de los textos a los cuales se aplica).

Las propuestas actuales abarcan la codificación morfosintáctica y de relaciones gramaticales de dependencias. Es importante señalar que el multilingüismo de las aplicaciones ha de ser contemplado para las que se prevean en el Estado español, y para la Unión Europea.

Además de las especificaciones para la información lingüística (morfosintáctica, dependencias, semántica, etc.), la reutilización de datos y procesadores tienen que fijar elección de otros detalles importantes para asegurar la interoperabilidad. El grupo de trabajo más activo actualmente en la definición de interoperabilidad para encadenar módulos en procesos de trabajo es CLARIN[54]:

- Codificación de caracteres (utf8)
- Segmentación del texto
- Lenguaje de etiquetas para codificar la información de anotación: XML, RD, tabular verticalizado, etc.
- Código de lengua
- Metadatos de descripción de recursos
- Formato para codificación de corpus (cuando se incluye más info que el texto plano): los más utilizados XCES y TEI
- Formato para codificación de memorias de traducción (empaquetado): se han implantado ya TMX y XLIFF
- Formato para codificación de diccionarios y terminología: LMF y TBX, son los de mayor implantación

Con respecto a la gobernanza de las infraestructuras lingüísticas es importante considerar la capacidad que tendrá la estructura propuesta para imponer, difundir, y fomentar la utilización de estos estándares, o en su defecto la capacidad de trabajo para desarrollar conversores que puedan realizar el trabajo de estandarización necesario para incluir un recurso nuevo. En las infraestructuras lingüísticas descritas anteriormente la gestión de los estándares se describe en la siguiente Tabla 1.

[54] http://clarin.ids-mannheim.de/standards/

	Catálogo y Metadatos	Interoperabilidad datos y procesadores	Catálogo descarga o API para uso	Evaluación y validación
CLARIN	sí	sí	sí	sí
META-SHARE	sí	como ecosistema	sí	no
LDC	no	no	parcialmente	sí
ELRA/ELDA	sí	no	no	sí
CREL	no	sí	no	no

Tabla 1: Gestión de estándares y otras características de las infraestructuras lingüísticas de los antecedentes

En cualquier caso, es importante introducir en el grupo de planificación estratégica una tarea relacionada con la vigilancia de estándares, especialmente en la industria e introducir en los grupos de trabajo y de dirección expertos que estén ya relacionados con los grupos internacionales de estandarización (aunque no está contemplado como área de trabajo por la institución europea competente, ETSI[55]). Para una revisión completa y referencias sobre los estándares aplicables se puede consultar el capítulo 5 del *Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital*[56].

6 CUESTIONES A TENER EN CUENTA PARA LA GOBERNANZA

A continuación, se propone una lista de cuestiones que deben tenerse en cuenta para diseñar la gobernanza de las infraestructuras lingüísticas: la participación, los procedimientos de toma de decisiones y de coordinación entre los participantes y con otras iniciativas que asegurarán la viabilidad del proyecto, su estabilidad y, muy especialmente, su sostenibilidad (preservación y crecimiento).

El estatus jurídico propio de las infraestructuras revisadas varía, pero es remarcable que en el ámbito español tanto el BSC como el IRTA se constituyeron como entidades de derecho público, consorcios, con personalidad jurídica propia, adscritas a departamentos relacionados de la administración estatal y autonómica. En las demás, la fórmula para que la infraestructura pueda ejercer actividad contractual y de administración de la infraestructura ha sido nombrar una entidad gestora, con personalidad jurídica propia.

[55] http://www.etsi.org
[56] http://www.agendadigital.gob.es/planes-actuaciones/tecnologias-lenguaje/Bibliotecaimpulsotecnologiaslenguaje/Material%20complementario/Informe-Tecnologias-Lenguaje-Espana.pdf

De las infraestructuras (o iniciativas similares) que específicamente abarcan recursos lingüísticos (datos y procesadores) hemos también estudiado el modelo de negocio o la forma de financiación y que resumimos en la siguiente tabla, junto con las características mencionadas antes.

Infraestructura/similar	Entidad gestora para operaciones	Entidad jurídica propia	Órgano máx. de gobierno	Financiación
CLARIN	Universidad de Utrecht, NL	European Research Infrastructure Consortium[57]	Asamblea General	financiación pública
META-SHARE	varias universidades y OPIs	no, depende de META-NET[58]	Junta META-NET	mantenimiento colaborativo y financiación pública puntual
ELRA	ELDA (empresa)	sí	Asamblea General	subscripción y pagos por distribución
LDC	Universidad de Pennsylvania	no	no	subscripción y pagos por distribución
BSC	--	Consorcio Entidad de derecho público	Consejo Rector	financiación pública
IRTA	--	Consorcio Entidad de derecho público[59]	Consejo de Administración	financiación pública y pagos por contratos y servicios
CREL	Institut d'Estudis Catalans	no	Consejo de dirección	financiación pública
VISC++	AQUAS (entidad de derecho público)	no	no	financiación pública

Tabla 2: Características de las infraestructuras lingüísticas de los antecedentes y modelo de negocio

[57] Being considered as an international organisation within the meaning of the directive on public procurement (Directive 2004/18/EC and Directive 2014/24/EC)

[58] META-TRUST AISBL (Association internationale sans but lucratif) is an international not-for-profit organisation based on Belgian law and located in Antwerp, Belgium.

[59] http://www.irta.cat/es-ES/LIrta/LleiDeLIrta/Paginas/default.aspx

La financiación pública de nuestras infraestructuras no está asegurada más allá de la duración del Plan TL. Existen dos esquemas posibles para su mantenimiento:

- Progresiva autofinanciación de la infraestructura mediante venta de licencias y cuotas por uso.
- Aportación colectiva de terceros que logran su financiación individualmente.

En España, la creación de recursos reutilizables hasta este momento se ha desarrollado, como se ha explicado en la sección 3.1.1, gracias al hecho de que la producción y distribución de recursos formaba parte de la misión del organismo proveedor bien como entidad normalizadora o como organismo dedicado a la investigación y la transferencia. Los primeros contaban con una dotación específica (organismos con capacidad normalizadora) y tenían capacidad para acceder a financiación de terceros (pública o privada), que es la única opción para los segundos.

El coste también puede ser cubierto por los usuarios, en un esquema de autofinanciación de la infraestructura. Es el modelo de distribuidores de productos de ELDA y LDC, los pagos por licencia se reparten entre el proveedor y el distribuidor, aunque en ambos casos acceden también a financiación pública en calidad de organismos especializados en la producción de recursos.

Por tanto, se deberá tener en cuenta que la implicación de proveedores y usuarios es lo que puede llegar a sustentar el mantenimiento y el crecimiento de la infraestructura creada por el Plan TL. La gobernanza de la infraestructura tendría que dotarse de canales de comunicación con ambos para implicarlos desde el inicio del diseño de su modus operandi.

Como ya se ha mencionado, la mayoría de los antecedentes revisados han optado por un claro desdoblamiento entre los mecanismos para garantizar las operaciones de los que se encargan de la planificación estratégica. Los tratamos ahora de forma separada.

6.1 CUESTIONES EN LA PLANIFICACIÓN ESTRATÉGICA

La gobernanza tiene que asegurar la comunicación entre los actores relevantes, el control de las agencias financiadoras con respecto al cumplimiento de los objetivos, y también la implicación de aquellos actores que van a ser la clave del mantenimiento y crecimiento de la infraestructura después del apoyo inicial de Plan TL. La primera cuestión a considerar es la participación en las tareas de planificación estratégica.

Partimos de una agrupación de los agentes revisados en la sección 3, en tres colegios, listados a continuación. Cabe preguntarse por la participación en la planificación estratégica de representantes los tres.

1. Agencias financiadoras
2. Proveedores de datos/procesadores, de servicios (almacenamiento, ejecución de procesos, etc.)
3. Usuarios

Agencias financiadoras

Las agencias financiadoras están presentes en los órganos de planificación estratégica de todas las infraestructuras revisadas en este documento que cuentan con financiación pública. Hay un aspecto en el que las infraestructuras lingüísticas del Plan TL van a tener una característica específica. Para las infraestructuras de las lenguas co-oficiales, las competencias sobre las cuales están cedidas a las correspondientes Comunidades Autónomas, la participación de las agencias financiadoras correspondientes debe ser tenida en cuenta.

En CLARIN, por ejemplo, las agencias son los únicos miembros de su Asamblea General con voto, y están asistidos por un consejo científico asesor y la Junta de Directores que es la responsable de la gestión operativa. En el BSC-CNS el Consejo Rector está formado por representantes de las agencias financiadoras y cuenta también con un consejo científico y una Junta de Directores que lo asesora. La diferencia con CLARIN es que en ésta al ser una infraestructura compuesta por diferentes centros de diferentes países la Junta de Directores engloba a todos los directores de centros estatales, mientras que en el BSC-CNS, los directores que pertenecen a la Junta son nombrados por el mismo Consejo Rector. El IRTA tiene como máximo órgano de gobierno un Consejo de Administración con representantes de las agencias financiadoras, pero también con vocales en representación de los diferentes colectivos implicados en la actividad, por ejemplo, expertos elegidos por una de las agencias financiadoras y un representante electo del personal del Instituto.

Proveedores

En cuanto a la participación de los proveedores tomamos primero a los proveedores de datos (listados en la sección 3.1). Por un lado, podemos separarlos entre aquellos que tienen como misión la provisión de datos a terceros, de los que no, con lo que las editoriales y publicaciones no serían consideradas para participar en la planificación. Esta misión del organismo proveedor puede ser crucial en el modelo

de sostenibilidad de las infraestructuras, ya que acabada la financiación del Plan TL seguirán cumpliendo su misión y pueden ser la clave de la sostenibilidad.

Además, puede parecer necesaria la distinción entre aquellos proveedores que aportan valor añadido a los datos (universidades, OPIs y organismos de normalización), de los que solamente son depositarios de datos crudos: bibliotecas y centros de datos. Tienen en común que ambos tipos tienen como misión suministrar datos a terceros, y que ambos tipos son proveedores con los que se puede establecer una relación contractual, de adquisición o licencias, de todos o parte de sus materiales. Hay que tener en cuenta, no obstante, que los datos crudos son (en su gran mayoría) de terceros, mientras que los datos de valor añadido que enriquecen el texto crudo pertenecen al proveedor y por tanto se han de tener en cuenta desde el punto de vista de los derechos de copyright. En este sentido, forman dos grupos diferentes. En lo que respecta a los proveedores de procesadores, se pueden asimilar a los proveedores de datos con valor añadido: se pueden establecer relaciones contractuales y son una parte importante de la sostenibilidad de las infraestructuras. En cualquier caso los unimos en lo que sigue bajo el título de proveedores.

La participación de los proveedores en la planificación estratégica puede provocar conflictos de intereses, en especial si se incluye a solamente algún proveedor (o algunos) que tienen de forma exclusiva información y voz en las decisiones sobre contratos de adquisición o producción de recursos. Incluirlos de forma colegiada (por ejemplo, un representante de las redes/asociaciones existentes) puede ayudar a mitigar el problema. En España, como se ha visto en la sección 3.1.1, los proveedores están asociados en la SEPLN y en redes como RETELE y TIMM cuyos representantes podrían actuar como interlocutores.

En los antecedentes vistos hay diferentes modelos. En CLARIN los Centros Nacionales son los proveedores (estos incluyen bibliotecas, centros de datos, dependiendo del país), pero éstos solamente participan en la gestión de operaciones y no en la planificación estratégica que lleva la Asamblea General. En META-SHARE están representados en la junta, pero hay que tener en cuenta que META-SHARE es una iniciativa colaborativa donde el mantenimiento de la infraestructura no tiene una agencia financiadora y los proveedores mismos son su motor. En ELRA-ELDA, al igual que en LDC, no hay distinción entre proveedores y usuarios: se trata de asociaciones de proveedores y usuarios en las que, además se cuenta con los pagos por licencias, aunque la relación de ELRA en relación con su empresa ELDA puede plantear también conflictos con los miembros proveedores. En cualquier caso se engloban todos como "miembros" y los beneficios materiales de ser miembros son descuentos en el precio y servicios de asesoría para la negociación de derechos y licencias. Los miembros del Consejo,

órgano de planificación, los son por votación entre los miembros de la Asamblea. En este sentido, es interesante ver que ELRA cuenta con un Comité de ética que resuelve los posibles conflictos de intereses. Por último, en IRTA es el gobierno de la infraestructura el que participa en la planificación estratégica de los centros consorciados que son los proveedores de servicios. En el Consejo de Administración los proveedores no están representados.

Usuarios

En las infraestructuras lingüísticas del Plan TL, los usuarios objetivo son las empresas, especialmente pymes y start-ups. En la mayoría de las infraestructuras consultadas la participación de usuarios se resuelve a través de la participación de expertos en los comités que asesoran la planificación estratégica o las operaciones. No obstante, es importante señalar que, debido a que se ha constatado que la mayoría de infraestructuras no acaban de tener el número de usuarios necesarios para su sostenibilidad, las recomendaciones[60] de diferentes grupos de reflexión sobre e-infraestructuras relevantes insisten en la necesidad de implicarlos en la planificación estratégica para que sus demandas y necesidades sean el verdadero objeto de la planificación.

Su participación puede también ser de forma colegiada, con representantes de asociaciones o clústeres empresariales del sector, actualmente los mencionados en la sección 3.4 de este documento.

Relación entre participantes

La participación de la agencia financiadora en el comité de máximo gobierno, durante el período del Plan TL pero también después, parece inexcusable.

La participación de proveedores y usuarios es recomendable para hacer posible la comunicación entre ellos. En las tareas de planificación estratégica señaladas en la sección 2.2, el grupo de gobierno ha de decidir sobre el plan estratégico, encargo de nuevas infraestructuras, etc. y una planificación estratégica conjunta para implicar a los productores en el mantenimiento de las infraestructuras e implicar a los usuarios para que éstos encuentren servicios para los que estén dispuestos a pagar cuotas o licencias, por ejemplo.

[60] http://e-irg.eu/ca/recommendations **:** *"These organisations should establish a clear separation between responsibilities for strategy setting and community building, operations, and innovation. Working with the user communities, they should strive to establish the e-Infrastructure umbrella forum for strategy setting in Europe, with sufficient user participation for community building, high-level strategy and coordination for the entire e-Infrastructure, with -again- a clear separation from operational responsibilities."*

Además, la toma de decisiones estratégicas es especialmente importante, como ya se ha anticipado, en las cuestiones que tienen relación con las especificaciones, o estándares, de interoperabilidad. La selección de estándares es una tarea técnica que puede llevarse a cabo por un grupo de trabajo específico tanto en la planificación estratégica como en las operaciones, pero las decisiones sobre la adopción han de ser vinculantes y afectaran todo el plan, por tanto, requieren del consenso entre proveedores y usuarios.

Parece recomendable, pues, la participación de proveedores y usuarios en la planificación y que en ambos casos haya una participación de representantes de estos colectivos, que deberían ser representantes elegidos por las asociaciones o clústeres de ambos colectivos. Como se ha mencionado antes los proveedores están asociados en la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) y existen actualmente tres clústeres relacionados con las industrias de las tecnologías de la lengua: Clustelingua, Langune y Plataforma del español.

6.2 CUESTIONES EN LAS OPERACIONES

En los antecedentes de la sección 4, se pueden distinguir dos modelos básicos de la organización de la gestión de las operaciones. El de entidad gestora y el distribuido. Pasamos ahora a describir cada uno de ellos en relación con las operaciones que en la sección 2.1 hemos clasificado en tres grandes grupos: adquisición/producción de recursos, distribución de los mismos y difusión-formación.

Modelo Entidad Gestora (ELRA/ELDA, LDC, VISC+), un organismo con entidad jurídica propia se encarga, por contrato, de las operaciones respecto a una colección de recursos y servicios que le son encomendados. En cuanto a las operaciones de ***adquisición/producción***, la entidad gestora es la encargada de gestionar una lista de proveedores con los que establece relaciones contractuales (para la adquisición, distribución, transformación de recursos), aplica los criterios de calidad establecidos llevando a cabo la evaluación de los mismos. La entidad gestora puede subcontratar la producción de nuevos recursos o desarrollarlos internamente.

En cuanto a la ***distribución,*** la entidad gestora mantiene el catálogo de recursos (uno propio o uno de alcance más amplio como META-SHARE), encargándose de la gestión de metadatos, y cubre la gestión y mantenimiento del repositorio o almacén de datos y procesadores. Por último, la operativa de ***difusión y formación*** recae en esta entidad que se hace responsable, por encargo, de la misma.

Modelo distribuido (CLARIN, META-SHARE, CREL, IRTA). Un número de centros son invitados a sindicar sus recursos y se les encarga, por contrato, llevar a cabo las operaciones dictadas por la máxima

autoridad de gobierno de la infraestructura. Se definen objetivos comunes o específicos para cada grupo de operaciones (distribución, adquisición y difusión-formación) que cada centro persigue de forma independiente según le ha sido encargado. El número y tarea de los centros lo define el máximo gobierno de las infraestructuras. CLARIN cuenta, además, con una Junta de directores, respaldada por una oficina, que también sirve al Director ejecutivo y al Director técnico.

En cuanto a la ***distribución***, una vez sindicados los recursos en un catálogo común, como el propuesto por META-SHARE o el VLO de CLARIN puede ser utilizado como catálogo central, donde cada nodo participante en las infraestructuras puede dar acceso a los recursos que tiene almacenados en cualquiera de los repositorios institucionales. Los nodos se encargan de las tareas de producción y gestión de metadatos. Es importante tener en cuenta la gestión del software del catálogo que normalmente se delega en un Director Técnico (CLARIN) o Comité de Coordinación (META-SHARE).

En cuanto a la ***adquisición/producción***, los centros pueden recibir de la autoridad de gobierno el encargo de producir nuevos recursos, o pueden desarrollar internamente nuevos recursos que propongan para ser integrados, previa evaluación, mediante adquisición o licencia. La *evaluación* y acreditación de los recursos puede encargarse al mismo centro productor, que debe seguir una metodología o criterios estándar o, quizá, la utilización de determinada plataforma de evaluación, o crearse una comisión específica en los órganos de gestión para decidir sobre la calidad de nuevos recursos.

En cuanto a la ***difusión y formación***, puede formar parte del contrato con los nodos, pero en la mayoría de casos es parte de las funciones del Director ejecutivo, como en CLARIN, o se delega esta función en otro organismo que tenga como misión la difusión y formación en el ámbito. META-SHARE, por ejemplo, delega esta función en la red de excelencia META-NET.

Al revisar las cuestiones de la gobernanza de la operativa, la garantía de sostenibilidad ha de estar desde el principio también en su diseño. Mantenimiento de las actividades y servicios de la infraestructura más allá del tiempo de la construcción en la que se recibe financiación específica, y mantenimiento (evolución) de los recursos depositados.

7 RECOMENDACIONES

Las recomendaciones para el modelo de gobernanza de las infraestructuras lingüísticas se estructuran en referencia a tres dimensiones: (i) principios del modelo, (ii) los órganos de la gobernanza y (iii) los procedimientos que median la toma de decisiones.

7.1 PRINCIPIOS SUBYACENTES AL MODELO DE GOBERNANZA

Las siguientes recomendaciones tienen en cuenta en primer lugar la **viabilidad** del proyecto, especialmente en el tiempo. La ventana que ofrece el Plan TL 2015-2020 es de cuatro años en el momento de presentar este informe.

En este período, que llamaremos *fase de construcción,* se deben cumplir los siguientes objetivos.

1. Que se cree una entidad que relacione los agentes necesarios, y que esta se dote de canales de comunicación y colaboración.
2. Que se pueda reunir un número crítico de recursos que sean interoperables en el menor tiempo posible para que la entidad inicie su actividad y contribuya a cumplir los objetivos del Plan TL.

En segundo lugar, se prioriza la **sostenibilidad** de las infraestructuras creadas. Estas han ser mantenidas después del período de financiación del Plan TL. El mantenimiento de las infraestructuras se refiere a que se mantengan operativas (servicios como el catálogo, el depósito, etc.) y a que se mantenga el esfuerzo en aumentar la cobertura de las mismas con más datos de valor añadido (nuevos corpus anotados, más niveles de anotación, etc.), más procesadores y actualización constante en ambos casos de los ya constituidos.

La duración limitada de la fase de construcción hace recomendable que el principio a seguir sea que la nueva entidad reúna y potencie los activos ya existentes:

- Reunir en la entidad a los agentes que puedan aportar recursos porque la creación de recursos y su distribución esté en su misión.
- Reunir actividades que ya están en marcha, aunque de forma dispersa.

La sostenibilidad ha de ser un compromiso fundacional de la infraestructura. Tanto la lengua oficial como las cooficiales cuentan con instituciones que elaboran recursos, la mayoría datos, pero también procesadores. Las hemos identificado en la sección 3.1.2 en el epígrafe "organismos de normalización lingüística". Son entidades de derecho público con la misión producir materiales que sean el referente de uso para una lengua. Desde este punto de vista, su participación en la creación y el mantenimiento de las infraestructuras supondría para ellas el reconocimiento oficial de un papel que ya están asumiendo: suministrar materiales que sean un referente en cuanto al uso de la lengua también en las

aplicaciones tecnológicas. Estas instituciones deben estar implicadas en la planificación de la infraestructura y elaborar conjuntamente con las agencias financiadoras una estrategia de sostenibilidad.

Al mismo tiempo, se ha de garantizar la comunicación entre todos los agentes y la transparencia de la planificación y las operaciones para conseguir un efecto colaborativo. La dotación económica es también limitada y, como ya se ha comentado, el objetivo de igualar en recursos a los disponibles para el inglés, en continuo crecimiento, hace recomendable no despreciar ninguna aportación.

La sostenibilidad también ha de ser adaptativa: las circunstancias pueden cambiar, pero se ha de mantener el esfuerzo. No solamente hay que pensar en inversión directa, también en la creación de un ecosistema que estructure la identificación de recursos y su reutilización. En este sentido juegan un papel importante las administraciones públicas que pueden aportar sus datos. Las empresas, que son los usuarios, juegan también un papel en la sostenibilidad ya que ha de ser la demanda y el retorno de la explotación de las infraestructuras las que creen un modelo de autofinanciación. Experiencias como LDC y ELDA parecen confirmar que es posible la sostenibilidad en un modelo mixto de subvenciones o contratos captados en convocatorias públicas y el pago por parte de empresas que comercializan productos que consumen recursos lingüísticos.

7.2 PLANIFICACIÓN. ÓRGANOS DE LA GOBERNANZA

Desde acuerdo con los dos principios y consideraciones anteriores, se propone estructurar la gobernanza mediante una entidad, en concreto una fundación/consorcio[61], que reúna a las agencias financiadoras y a las entidades que tienen como misión la generación y mantenimiento de recursos lingüísticos de valor añadido para formar el órgano máximo de gobierno. Esta entidad garantizará la coordinación en los desarrollos de las diferentes lenguas de España, garantizando la interoperabilidad necesaria para un mercado interno multilingüe.

Las entidades normalizadoras, en particular las academias de la lengua son los organismos únicos por lengua, cuya misión es desarrollar y proporcionar recursos lingüísticos a terceros. Se les ha de proponer ahora que lo hagan también para las aplicaciones tecnológicas que consumen información lingüística. Consecuentemente se les ha de considerar una pieza clave de las infraestructuras y

[61] *Queda pendiente determinar la adecuación: Consorcio es la entidad cuya finalidad es la realización o prestación de servicios de forma asociativa sobre asuntos de interés común; Fundación es la organización constituida sin ánimo de lucro que dedica su patrimonio público a fines de interés general.*

servicios también en los productos de las tecnologías de la lengua. Las academias de la lengua actúan también como organismos que conocen y colaboran con otros proveedores para las lenguas de su incumbencia. Junto con las agencias financiadoras relevantes (por cuestiones de competencias de las comunidades autónomas con lengua cooficial) formarían el núcleo del órgano de máximo gobierno de las infraestructuras. Este órgano (el patronato, si se trata de una fundación) garantizaría la planificación estratégica y delegaría las operaciones en entidades gestoras de las infraestructuras de cada lengua. La entidad se encargaría de la federación de recursos.

Esta entidad tendrá el objetivo de construir y mantener las infraestructuras lingüísticas para las lenguas de España oficial y cooficiales.

El Patronato/Consejo de Administración

Tendrá las funciones y responsabilidades siguientes:

- Servirá de mecanismo de coordinación, colaboración, intercambio de experiencias y ayuda mutua entre los órganos competentes de la Administración General del Estado, y de otras Administraciones, en particular, de las Comunidades Autónomas con lengua cooficial, para organizar de forma coordinada y consensuada la construcción de la infraestructura de todas ellas.
- Evaluar la situación inicial para las distintas lenguas e infraestructuras para planificar en detalle las necesidades futuras.
- Aprobar la planificación operativa de la construcción de la infraestructura. Debe detallar las actuaciones a realizar conjunta o individualmente por las partes implicadas, cómo se evaluará el progreso y el logro de objetivos, etc.
- Evaluará periódicamente el avance de la planificación y su impacto para evaluar su cumplimiento y planificar las siguientes fases.
- Propondrá modificaciones a la planificación de modo adaptativo a las circunstancias.
- Nombrará y cesará a los miembros de las comisiones delegadas.
- Facilitará la coordinación y coordinación con infraestructuras parecidas europeas e internacionales.

Estará compuesto por representantes de las agencias financiadoras y las academias de la lengua de las lenguas de España oficial y cooficiales.

Fundadores

1. El secretario de Estado para la Sociedad de la Información y la Agenda Digital.
2. El que designe la Generalitat de Catalunya.
3. El que designe el Gobierno Vasco
4. El que designe la Xunta de Galicia
5. El director/a de la Real Academia Española
6. El director/a de la Real Academia de la Lengua Gallega
7. El director/a de la Real Academia de la Lengua Vasca
8. El director/a del Institut d'Estudis Catalans

Vocales (no necesariamente relacionados con los fundadores)

- El director/a ejecutivo
- Dos vocales, a propuesta de la SESIAD
- Un vocal a propuesta de cada uno de los demás patronos/fundadores

El órgano de gobierno de esta entidad estará asesorado por un **Comité Asesor** internacional que garantizará (i) las conexiones con las infraestructuras que, como hemos visto en la sección 4, ya existen a nivel europeo, (ii) con empresas especialmente relevantes del área y (iii) con expertos de reconocido prestigio tanto nacionales como internacionales.

El órgano de gobierno nombrará un Director/a Ejecutivo, que actuará de secretario, y que gestionará la entidad creada, la ejecución de la planificación y el control de las operaciones comunes (ver más abajo) asistido por una secretaría de la entidad.

7.3 TOMA DE DECISIONES

Las decisiones del máximo órgano de gobierno se tomarán por mayoría cualificada de dos tercios.

Típicamente las decisiones del máximo órgano de gobierno estarán motivadas, además de por las cuestiones de la ejecución, por las propuestas de las siguientes comisiones delegadas.

Figura 10: Diagrama de la relación entre los órganos que participan en la planificación de las infraestructuras lingüísticas del Plan TL. Recomendaciones.

7.4 COMISIONES DELEGADAS

La entidad deberá establecer estatutariamente comisiones delegadas, como mínimo, en las siguientes áreas específicas: Comisión de Planificación Estratégica; Comisión de Evaluación, Adquisición y nuevos desarrollos; Comisión de Estándares y Licencias y Comisión de Servicios. Nombrará, también de entre los vocales, presidentes de las comisiones. El Director Ejecutivo actuará de secretario. Su composición podrá variar a lo largo de la realización del Plan. En cada comisión deberán estar representados al menos los proveedores de datos, los proveedores de procesadores y los usuarios del ámbito industrial de las tecnologías del lenguaje. Los miembros, a propuesta del presidente, serán nombrados por el órgano de gobierno a título personal o como representantes de los colectivos. Estas comisiones se convierten así en el foro de participación de estos colectivos cuya opinión y demandas son cruciales para la planificación e implementación correcta de la misión de la infraestructura. Cada comisión deberá tener un mínimo de cuatro miembros, además del presidente y el secretario, y un máximo de doce.

Estas comisiones recibirán el mandato del comité máximo de gobierno de elaborar propuestas y recomendaciones que acometan las tareas de planificación específicas mencionadas en el Plan TL, las listadas en la sección 2.2, así como la participación de un representante en comités, reuniones y foros internacionales etc. que sean considerados necesarios para garantizar la correcta ejecución de las medidas del eje 1 del Plan TL. Serán funciones de todas las comisiones delegadas las siguientes:

- Asesorar al órgano de máximo gobierno

- Servir de mecanismo de interlocución entre los actores implicados en la construcción y posterior sostenibilidad de la infraestructura.
- Colaborar en la difusión de la infraestructura en sus respectivos sectores.

Tratamos a continuación la misión específica encomendada a cada comisión.

7.4.1 COMISIÓN DE PLANIFICACIÓN ESTRATÉGICA

Tendrá la misión de elaborar una propuesta de plan estratégico que asegure la sostenibilidad después de la fase de construcción financiada por el Plan TL. Debe proponer un modelo de negocio que abarque la sostenibilidad y la escalabilidad de las infraestructuras.

7.4.2 COMISIÓN DE EVALUACIÓN, ADQUISICIÓN Y NUEVOS DESARROLLOS

Tendrá la misión de planificar la colección y adquisición de infraestructuras durante la fase de construcción, identificando métodos de evaluación de recursos ya existentes (datos y procesadores), especificando las necesidades que manifiesten en la comisión proveedores y usuarios, y proponiendo un plan de nuevos desarrollos y posibles proveedores.

Velará por la creación de un inventario de infraestructuras existentes y lo irá actualizando durante la vida del Plan TL teniendo en cuenta las infraestructuras que sean resultado de las acciones de la entidad o de obras de terceros y que puedan ser incorporadas. La participación de usuarios y proveedores es de suma importancia para estudiar y vigilar la relación entre oferta y demanda y proponer acciones de adquisición y nuevos desarrollos.

La comisión elaborará un plan de desarrollo con carácter anual que constituirá el programa a ejecutar por los diferentes nodos, estableciendo las características técnicas en colaboración con los nodos de lengua para que éstos lo implementen. El órgano de gobierno aprobará este plan y dotarlo de la financiación necesaria para que pueda implementarse.

Típicamente, esta comisión debe elaborar una propuesta de programa de desarrollo anual que una vez aprobado por el órgano de gobierno constituya el mandato para los nodos de lengua (ver operaciones).

Esta comisión también será la encargada de planificar campañas periódicas de evaluación de datos y procesadores, proponiendo la participación de los nodos de la infraestructura en la organización de las mismas.

7.4.3 COMISIÓN DE ESTÁNDARES Y LICENCIAS

Tendrá la misión de seleccionar las normas técnicas de interoperabilidad, los modelos de licencias y los mecanismos de protección de datos personales adecuados que deberán implementarse en el despliegue de las infraestructuras. Las propuestas de esta comisión deberán tener en cuenta los intereses de los usuarios y de los proveedores, así como que se cubran las necesidades de la lengua oficial y las cooficiales en sus decisiones.

7.4.4 COMISIÓN DE SERVICIOS

Tendrá la misión de proponer y vigilar los medios de acceso público a las infraestructuras y un catálogo de servicios que promuevan y fomenten la reutilización de las mismas por parte de los usuarios, que estarán debidamente representados en la comisión.

7.5 OPERACIONES. MODELOS Y ORGANIZACIÓN

Dadas las características de la entidad creada, que reúne cuatro instituciones cada una especializada en una lengua particular, proponemos un modelo distribuido para la organización las operaciones que implementarán las decisiones del órgano de máximo gobierno, al menos en el primer nivel. Cada academia se convierte así en una entidad gestora para las operaciones de la infraestructura como nodo para la lengua de su incumbencia.

Figura 11: Diagrama de la relación entre los órganos que participan en las operaciones de las infraestructuras lingüísticas del Plan TL. Recomendaciones.

En tanto que entidades gestoras para las operaciones, las academias reciben el encargo llevar a cabo operaciones específicas (relacionadas con la lengua de su incumbencia) que pueden realizar de forma autónoma. Las funciones de los nodos de lengua, bajo la forma que finalmente tomen, son las que se deriven del cumplimiento de las operaciones listadas en la sección 2.1 según la planificación que haya sido acordada por el órgano de máximo gobierno.

En cuanto a las operaciones de adquisición, producción y evaluación, las academias en el cumplimiento de la planificación acordada en el órgano de máximo gobierno y con la financiación adjudicada, pueden, de acuerdo a sus circunstancias particulares, cumplir internamente el encargo, subcontratar el trabajo a proveedores, delegar a su vez en otra entidad parte o el total del encargo, u organizarse a su vez según un modelo distribuido constituyéndose en un nodo central de las infraestructuras para esa lengua que está compuesta de diferentes nodos cada uno con una especialización o papel en la infraestructura.

En cuanto a las operaciones de distribución y las de difusión y formación las academias, en tanto que entidades gestoras contribuirán (o harán contribuir) a las operaciones comunes, coordinadas por el Director Ejecutivo. Estas operaciones comunes tendrán como eje básico la sindicación de la información sus recursos y se definirán objetivos para cada grupo de operaciones. Con este catálogo se contribuirá a las operaciones comunes de distribución, mediante un catálogo común (a imagen de META-SHARE o el VLO de CLARIN) a la que todas las academias se comprometen a contribuir. El catálogo común es, no obstante, responsabilidad de la Dirección Ejecutiva.