

Reseña de Rockwell, G., & Passarotti, M. (2019). The Index Thomisticus as a Big Data Project. *Umanistica Digitale*, 3(5).


<https://doi.org/10.6092/issn.2532-8816/8575>

Reseña realizada por:

Cristina MUÑOZ

8422.munoz@gmail.com

American College of the Mediterranean

 <https://orcid.org/0000-0001-8816-2992>

Cita recomendada:

Muñoz, C. (2024). Reseña de Rockwell, G., & Passarotti, M. (2019). The Index Thomisticus as a Big Data Project. *Umanistica Digitale*, 3(5).. *Publicaciones de la Asociación Argentina de Humanidades Digitales*, 5, e66.
<https://doi.org/10.24215/27187470e066>

RECIBIDO: 28/06/2024 ACEPTADO: 12/12/2024

El artículo reseñado recoge el trabajo de los profesores Geoffrey Rockwell (University of Alberta, Canadá) y Marco Passarotti (Università Cattolica del Sacro Cuore, Italia) sobre el proyecto del *Index Thomisticus* (IT), el cual inició el sacerdote jesuita Roberto Busa (1913-2011) a comienzos de la década de 1950. El IT surge como una herramienta para realizar búsqueda lematizadas dentro de la obra completa de Santo Tomás de Aquino. El objetivo que perseguía Busa era lograr un método que permitiera registrar, identificar, recuperar y procesar de manera automática y replicable todas las ocurrencias de las 11 millones de palabras que componen la obra de Santo Tomás de Aquino.

Rockwell y Passarotti estructuran el artículo en tres partes: la primera contextualiza el periodo en el que se desarrolló el proyecto de Busa; la segunda trata de los aspectos específicos del IT; y la tercera es un reflexión de los autores sobre el legado de Busa y el desarrollo de las Humanidades Digitales.

Rockwell y Passarotti comienzan señalando un problema recurrente en los proyectos digitales: la dependencia de un software. El software otorga la infraestructura necesaria para el análisis de los datos específicos al proyecto, por lo que, a través de él, se puede saber cuáles fueron las preguntas iniciales que motivaron una investigación concreta. Sin embargo, es frecuente que los proyectos tiendan a abandonarse cuando el software deja de funcionar, bien porque evoluciona, bien porque no cumple con su cometido. En este sentido, Rockwell y Passarotti abogan por ocuparse de la historiografía en la que se desarrollan los proyectos. Los autores afirman que el IT merece un especial reconocimiento no solo porque se encuentra a medio camino entre los proyectos de las humanidades tradicionales y las digitales, sino también porque consiguió establecer un método para trabajar con el lenguaje no estructurado cuando no había lo que hoy se considera un ordenador¹. El material que Busa utilizó para la creación de los archivos del proyecto también es muy diverso, ya que abarca documentación personal, como certificados académicos o documentos de identidad, documentación de índole profesional, como conferencias y papeles de preparación de seminarios, congresos y talleres; información logística, como reservas de hoteles, trenes, vuelos o facturas; artículos de prensa sobre el desarrollo del proyecto, correspondencia del ámbito personal y profesional con diferentes académicos y responsables administrativos y religiosos, material relativo a las fases del proyecto, y fotografías y publicaciones que Busa realizó en el tiempo del desarrollo del proyecto.

¹ El Index abarca un amplio periodo de tiempo que va desde principios de la década de 1950 a 2010, por lo que no es comparable la tecnología informática y el acceso a los ordenadores cuando se inició el proyecto con el desarrollo de los últimos años.

Gracias a estos materiales archivados, se puede conocer cómo se desarrolló el IT. Rockwell y Passarotti lo analizan desde cuatro puntos de vista: las comunicaciones, la concepción, el proceso y el contexto histórico.

La primera parte, las comunicaciones, fue el inicio del proyecto donde Busa se comunica con investigadores y estudiosos internacionales en busca de apoyo para su proyecto, con la idea de presentárselo a IBM², avalado por la comunidad internacional. No obstante, estas comunicaciones no tendrían sentido sin una previa concepción del proyecto.

En la búsqueda de ese apoyo entre sus colegas, Busa elaboró una serie de resúmenes de las fases del proyecto para que el resto de investigadores pudiera comprender de qué se trataba. Un ejemplo se encuentra en la Introducción del *Varia Specima Concordantium*, donde Busa resume las cinco fases necesarias para compilar la concordancia de la obra de Santo Tomás de Aquino y crear un índice de palabras procesado por máquinas:

- Transcripción del texto, desglosado en frases, en tarjetas separadas³.
- Multiplicación de las tarjetas (tantas como palabras haya en cada una de ella).
- Indicación de la entrada respectiva, esto es, del lema, en cada tarjeta.
- Selección y colocación en orden alfabético de todas las tarjetas según el lema y su cualidad puramente material.
- Composición tipográfica de las páginas para su publicación, tras la elaboración formal del orden alfabético de las palabras.

² Busa logró la colaboración de la multinacional IBM, que prestó las máquinas mecanográficas, los computadores electrónicos y los servicios tecnológicos necesarios para la ejecución del proyecto.

³ Las tarjetas perforadas eran láminas hechas de cartulina que contenía la información en forma de perforaciones según un código binario. Por otro lado, según los autores, Busa se refería con la palabra "frase" a una perícopa, es decir, una unidad de pensamiento coherente.

Pronto, en 1957, con la colaboración de Paul Tasman⁴, estas fases se redujeron a dos:

- Análisis del texto por parte de un investigador, marcándolo con instrucciones precisas para que se puedan perforar las tarjetas.
- Copia del texto por parte de un empleado del proyecto, utilizando una máquina de escribir especial que opera con una perforadora de tarjetas.

Con el paso del tiempo, Busa y Tasman se dieron cuenta de que su proyecto podía ir más allá de la concordancia, ya que empezaron a plantearse cómo se podía usar para el análisis lingüístico y la ingeniería del lenguaje.

Por otro lado, Rockwell y Passarotti hacen una llamada de atención al lector sobre el proceso de elaboración del IT y observan las dos innovaciones que este introduce: la representación del texto plano para el procesamiento computacional, el desarrollo de la tokenización para el procesamiento de los datos y la generación de varios tipos de índices.

Los autores no pierden de vista el contexto histórico en el que se desarrolló el proyecto y sin el cual no se entendería el IT. Rockwell y Passarotti enumeran una serie de situaciones específicas a las que se enfrentó Busa y que condicionaron su proyecto porque supusieron un mayor aporte tanto en recursos humanos como financieros. En primer lugar, Busa y su equipo se vieron en la obligación de transcribir la obra de Santo Tomás porque, debido a la época, no era posible descargarla y convertirla a texto plano, sino que los textos necesitaban ser perforados tras copiarlos manualmente para que las máquinas electromecánicas pudieran clasificar, replicar e imprimir en tarjetas perforadas la información registrada. Estas tarjetas, que funcionaban como portadoras de información, era la única tecnología estandarizada al inicio del proyecto. Las tarjetas del momento podían manipularse tanto mecánica

⁴ Tasman fue el ejecutivo de IBM que trabajó en el IT.

como manualmente, por lo que se creaban dos capas de información: la que se procesaba por máquinas y la que se procesaba por los colaboradores a través de las anotaciones de los académicos.

Tampoco había un mecanismo estándar para el procesamiento de textos; de hecho, las máquinas que procesaban las tarjetas perforadoras se programaban desde cero para cada proyecto. De acuerdo con Rockwell y Passarotti, esto tenía la ventaja de que Busa y su equipo no debían preocuparse por cómo formateaban los datos o por el software de las personas externas al proyecto que quisieran consultar la información porque las máquinas clasificadoras de las tarjetas electromecánicas no contaban con un sistema operativo.

En cuanto a la técnica de tokenización, Busa y Tasman desarrollaron dos tipos de tarjetas⁵: tarjetas de frases (*Phrase Cards*) y tarjetas de palabras (*Each Word Cards*). La técnica consistía en dividir la obra en fragmentos más pequeños, ya que en primer lugar, se usaban las tarjetas perforadas con las frases que habían sido codificadas (*Phrase Cards*) y luego, se procesaban para obtener cada palabra (*token*) en el texto (*Each Word Cards*). Estas últimas tarjetas describían los datos para la localización de las palabras en el texto. De este modo, la información que contenía era la siguiente:

- Referencia a la tarjeta de frase, es decir, su ubicación.
- Marca de referencia especial debido a la limitación del espacio en las tarjetas.
- La palabra en sí tal y como aparece en el texto.
- Orden numerado de las palabras en el texto.
- Primera letra de la palabra precedente.
- Primera letra de la siguiente palabra.

⁵ Estas tarjetas serían equivalentes a bases de datos. Como estaban vinculadas a través de la técnica que emplearon, se generaron dos índices que permitían realizar múltiples funciones, tales como contar palabras, clasificarlas, recuperar los textos que contenían determinadas palabras para construir una concordancia, etc.

- Número de Tarjeta de Forma, esto es, los tipos de palabras ordenados alfabéticamente.
- Número de Tarjeta de Entrada o encabezados de los tipos de palabras tras la lematización y la desambiguación.

Finalmente, los autores del artículo reflexionan sobre el presente y el futuro de las HD. Proponen que se vuelva al término Humanidades Computacionales para recalcar que el trabajo de investigación se realiza sobre los datos, ya que en la actualidad, cualquier tipo de trabajo implica en sí mismo un cierto grado de digitalización. Los datos son el nexo común entre el enfoque tradicional y digital de las Humanidades, por lo que no deberían estar diferenciadas, sino basadas en evidencias empíricas rigurosas y de calidad a través de la aplicación de métodos, técnicas, recursos y herramientas computacionales.

Rockwell y Passarotti defienden que los datos de cualquier investigación sean localizables, accesibles, interoperables y reutilizables:

- Deben tener un identificador global único y persistente (Localizable).
- Se deben recuperar mediante un protocolo de comunicación abierto, libre y estandarizado (Accesible).
- Deben utilizar un lenguaje formal, accesible, compartido y aplicable para la representación del conocimiento (Interoperable).
- Deben publicarse con una licencia de uso clara, accesible y con una procedencia detallada (Reutilizable).

Esta descripción de los datos es la que se propuso para los proyectos Horizon 2020 de la Unión Europea y recoge los principios en el tratamiento de los datos por parte de Busa: repetición, replicación, reproducción y reutilización, manteniendo, de esta forma, su legado.

A modo de conclusión, hay que señalar que el IT ha evolucionado en la creación de proyectos que continúan complementando el proyecto de Busa, pero también ha influenciado otros proyectos ajenos a él. Así, en

2005, se desarrolló la versión web del corpus⁶, donde se pueden observar grafos que representan las relaciones conceptuales que se dan en un plano semántico dentro de la obra de Santo Tomás. Un año más tarde, se anexaron las anotaciones morfosintácticas para crear un *treebank*, bajo la coordinación de uno de los autores del artículo, Passarotti⁷.

Por otro lado, el método de Busa influenció otros proyectos como la aplicación de la concordancia de Luhn conocida como *Key Word In Context* (KWIC).. También inspiró la creación del software Antconc⁸, cuya funciones para lista de palabras (*Word List*, todas las palabras del corpus con la frecuencia de aparición), concordancia (*Concordance*, búsqueda de una palabra en el corpus), token (cantidad de palabras y repetición en el corpus), *type* (cantidad de palabras diferentes), *collocates* (relación de un término concreto con otros; permite la observación de patrones) o anotaciones (*Annotations*, categoría gramatical de las palabras), se asemejan a la metodología de las tarjetas perforadas de Busa.

A día de hoy, el IT está incluido en la base de datos LiLa⁹, la cual es una fuente de recursos lingüísticos y herramientas de procesamiento del lenguaje natural para el latín, preservando así el trabajo iniciado por Busa.

⁶ Accesible desde: <https://www.corpusthomicum.org/>.

⁷ Accesible desde: <https://itreebank.marginalia.it/view/projet.php>.

⁸ Accesible desde: <https://www.laurenceanthony.net/software/antconc/>.

⁹ Accesible desde: <https://lila-erc.eu/#page-top>.