

# **Trabajo de Investigación**

## **Estado del arte de la investigación sobre wikis**

Emilio José Rodríguez Posada

Tutor: Juan Manuel Doderó Beardo

Universidad de Cádiz

Diciembre de 2012

# ¿Qué es un wiki?

- Un **wiki** es un **sitio web**
- Que permite a los visitantes:
  - Consultar los contenidos
  - Y también ¡**modificarlos!**
- El primer wiki se inventó en 1996
- Y en 2001 se fundó Wikipedia:
  - Wikipedia es el ejemplo más conocido
  - Pero **existen miles de wikis** por toda la red



Ward Cunningham, inventor del primer wiki.  
Foto: Carrigg Photography (CC BY-SA 3.0)

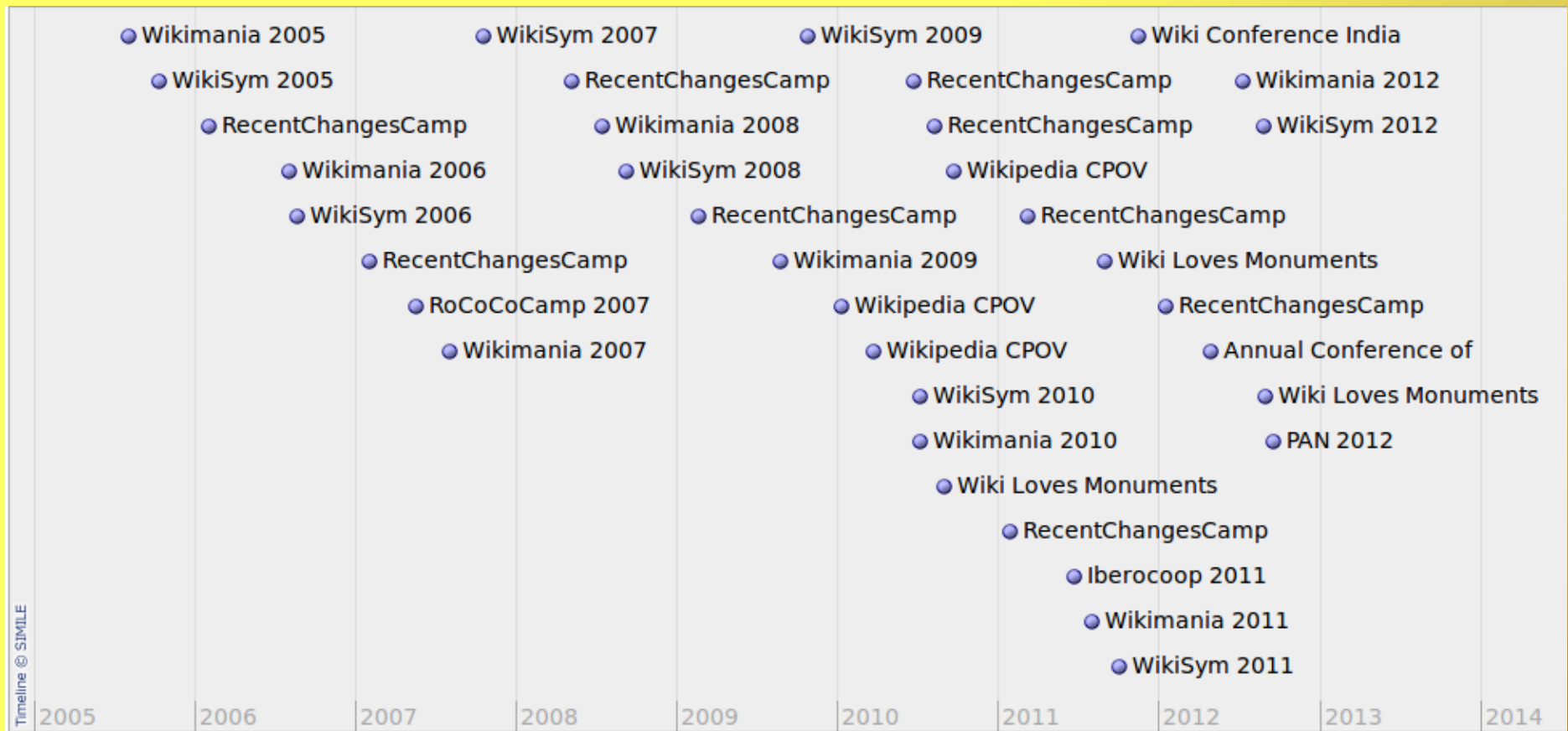
# ¿Por qué investigar los wikis?

- Los **usuarios** al participar en los wikis **generan muchos datos** de forma indirecta:
  - Estos datos se quedan guardados en el historial
  - Las conversaciones de los usuarios son visibles
  - Y también existen registros de:
    - Visitas, bloqueos, ataques, etc
- Todos estos datos son **públicos**:
  - Y generalmente tienen licencia libre (CC, GFDL, ...)
- De forma que se facilita acceder y analizar estos datos:
  - Se puede estudiar cómo funcionan, sus dinámicas, sus estructuras, el comportamiento de los usuarios, etc

# ¿Hay actividad investigadora? (I)

- Desde el año 2005 ha ido creciendo el interés por los wikis. Existen numerosos:
  - **Congresos:** WikiSym, CLEF/Pan Lab
  - **Workshops:** WikiAI, SemWiki, MathWikis
  - **Conferencias:** Wikimanía, WikiCon, SMWCon, Wiki Conference India, Wikipedia Academy, Wikipedia CPOV Conference
  - **Competiciones:** WikiViz (visualización datos)

# ¿Hay actividad investigadora? (II)



Se observa como cada año que pasa hay más eventos


# Mucha actividad = Muchos *papers*

- Ha habido **varios intentos de recopilar toda la literatura** sobre wikis. Pero han fallado por el camino:
  - Han resultado ser **incompletos**
  - Han sido **abandonados** con el tiempo y se han perdido los datos
  - No explotan la potencia de los metadatos de las publicaciones
    - Por ejemplo con web semántica
  - Y hay escasa o **nula generación de gráficas**, tablas, estadísticas, etc
- Algunos **ejemplos fallidos** son: grupos de Zotero o CiteULike, recopilaciones en webs personales, etc. Pero todo estaba muy disperso e incompleto.
- Para investigar sobre wikis **necesito conocer el estado del arte**:
  - **Esto me lleva a inventar mi propio sistema: WikiPapers**

<http://wikipapers.referata.com>

# ¿Por qué hacerlo en un wiki?

- Es fácilmente editable
  - No requiere registro
  - **Es colaborativo**
- Es enriquecido con **semántica**
  - Permite búsquedas, filtrados
  - Inferir conocimiento
- Generación de listados:
  - En base a metadatos
- Generación de **estadísticas**:
  - Gráficas
  - Tablas



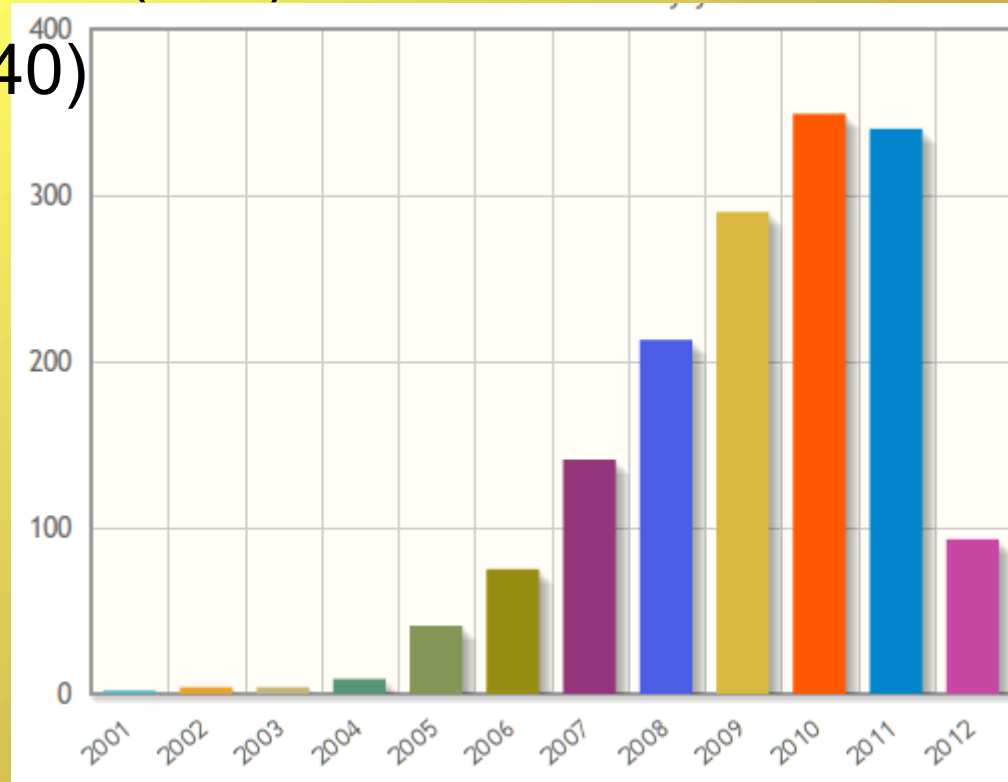
The screenshot shows the WikiPapers website interface. The main page features a navigation menu on the left, a search bar at the top right, and a central content area. The content area includes a welcome message, a list of publications, and a bar chart titled "Publications distribution by year". The bar chart shows the number of publications from 2001 to 2012, with a significant increase starting around 2008.

Year	Number of Publications
2001	0
2002	0
2003	0
2004	0
2005	0
2006	0
2007	0
2008	10
2009	20
2010	40
2011	130
2012	200

Captura de pantalla WikiPapers

# WikiPapers: estado actual

- El proyecto **WikiPapers** incluye por ahora:
  - Lista de **publicaciones** sobre wikis (+1700)
  - Lista de **conjuntos de datos** (+90)
  - Lista de **herramientas** (+40)
  - Lista de **autores** (+1000)
- También información de:
  - Revistas, eventos
  - Preguntas abiertas
  - Ejemplos
  - Y mucho más...



Número de publicaciones por año



# ¿Qué información se recoge? (I)

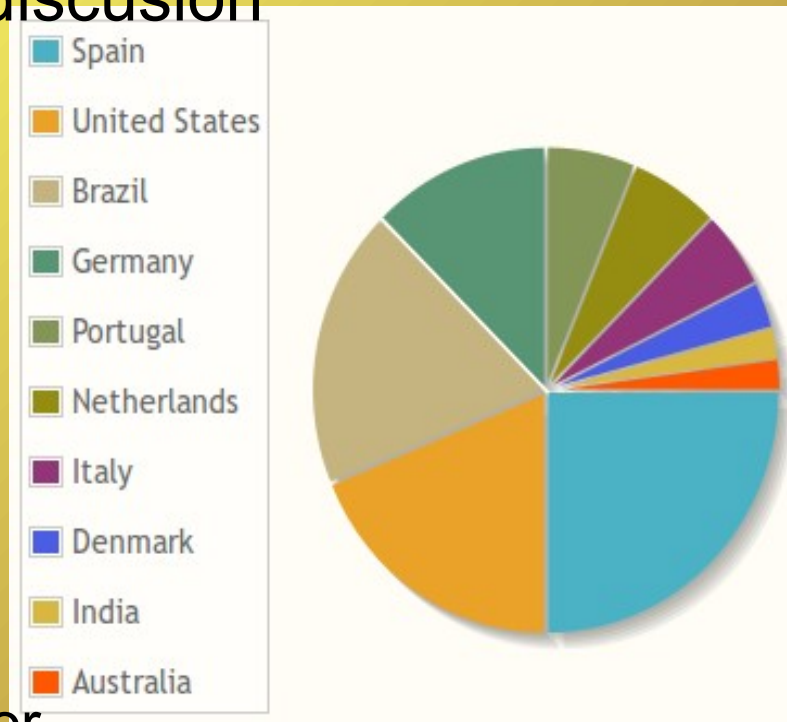
- Sobre las **publicaciones**:

- Título, autores, palabras clave, año, revista o congreso, DOI, idioma, licencia, enlaces al fichero y motores de búsqueda, abstract, las referencias que incluye, las citas que recibe y un espacio de discusión

- Sobre los **autores**:

- Nombre, afiliación, país
- Índice de coautores
- Página web
- **Estadísticas:**

- Número de publicaciones
- Y número de citas que recibe el autor



Distribución autores por país

# ¿Qué información se recoge? (II)



Emijrp My talk Site settings Admin links My preferences My watchlist My contributions Log out

Page Discussion View Edit Edit source History

## Collective memory building in Wikipedia: The case of North African uprisings

**Collective memory building in Wikipedia: The case of North African uprisings** is a 2011 conference paper written in English by Michela Ferron, Paolo Massa and published in WikiSym.

### Abstract [edit]

Since December 2010, a series of protests and uprisings have shocked North African countries such as Tunisia, Egypt, Libya, Syria, Yemen and more. In this paper, focusing mainly on the Egyptian revolution, we provide evidence of the intense edit activity occurred during these uprisings on the related Wikipedia pages. Thousands of people provided their contribution on the [content pages](#) and discussed improvements and disagreements on the associated [talk pages](#) as the traumatic events unfolded. We propose to interpret this phenomenon as a process of [collective memory](#) building and argue how on Wikipedia this can be studied empirically and [quantitatively](#) in real time. We explore and suggest possible directions for future research on collective memory formation of traumatic and [controversial](#) events in Wikipedia.

### References [edit]

This publication has 14 references. Only those references related to wikis are included here:

- "Fixing the floating gap: the online encyclopaedia Wikipedia as a global memory place"
- "Why do people write for Wikipedia? Incentives to contribute to open-content publishing"
- "Readers are not free-riders: reading as a form of participation on Wikipedia"
- "WikiChanges: exposing Wikipedia revision activity"
- "Studying collective memories in Wikipedia" (create it!) [\[search\]](#)
- "Exploring Linguistic Points of View of Wikipedia"

### Collective memory building in Wikipedia: The case of North African uprisings

**Author(s)** Michela Ferron, Paolo Massa

**Published in** WikiSym

**Date** 2011

wikipedia, web 2.0, collective memory, revolutions, traumatic

**Keyword(s)** events, Egypt, North Africa (**Extra:** egyptian revolution, arab spring, history, current events)

**Peer-reviewed?** Yes

**Language(s)** English

**License(s)** Unknown [\[+\]](#)

**Identifiers**

**ISBN** Unknown [\[+\]](#)

**DOI** Unknown [\[+\]](#)

**OCLC Number** Unknown [\[+\]](#)

**CiteUlike** Unknown [\[+\]](#)

Ejemplo de *paper* con metadatos a la derecha, abstract y referencias

[http://wikipapers.referata.com/wiki/Collective\\_memory\\_building\\_in\\_Wikipedia:\\_The\\_case\\_of\\_Nor](http://wikipapers.referata.com/wiki/Collective_memory_building_in_Wikipedia:_The_case_of_Nor)

# ¿Qué información se recoge? (III)

Emijrp My talk Site settings Admin links My preferences My watchlist My contributions Log out


Page **Discussion** View Edit Edit source History  Go Search

## Felipe Ortega


Felipe Ortega is an [author](#) from [Spain](#).

### Tools

Tool	Description
WikiXRay	WikiXRay is a robust and extensible software tool for an in-depth quantitative analysis of the whole Wikipedia project.



**Felipe Ortega**  
*(Alternative names for this author)*



Felipe Ortega

**Affiliation** Unknown [+]

**Country** Spain

Antonio J. Reinoso, Daniel Izquierdo-Cortazar, Gregorio Robles, Israel Herraiz, Jesús M. González-Barahona, Joaquín Rodríguez López, Joseph M. Reagle, Luca de Alfaro, **Rut Jesus**

**Co-authors**

**Website** <http://felipeortega.net/>

**Statistics**

**Authorship** Publications (10), datasets (0), tools (1)

**Citations** Total (14), average (1.4), median (0),

Navigation

- Main page
- Publications
- Keywords
- Authors
- Datasets
- Tools
- Events
- Journals
- Concepts
- Open questions
- Research areas
- Examples
- ...More lists...
- Random page

Create new...

- Publication
- Author
- Event
- Keyword

Ejemplo de autor con metadatos a la derecha, lista de herramientas, vídeo incrustado...  
[http://wikipapers.referata.com/wiki/Felipe\\_Ortega](http://wikipapers.referata.com/wiki/Felipe_Ortega)

# ¿Qué información se recoge? (IV)

- También hay información sobre:
  - **Conjuntos de datos (*Datasets*)**
    - Bases de datos de Wikipedia (son públicas)
    - Y también de otros wikis de Internet
  - **Herramientas**
    - *Frameworks* para extraer datos de wikis
    - Procesamiento de datos
    - Estadísticas y visualización
- Todos los **metadatos** de publicaciones, autores, *datasets* y herramientas son **exportables**:
  - En los formatos BibTeX, RDF, CSV y JSON

# Estado del arte (I)

- Se ha utilizado el siguiente procedimiento:
  - **Términos:** wiki, Wikipedia, Wiktionary, Wikibooks, Wikiversity, Wikiquote, Wikisource, Wikinews, Wikispecies, MediaWiki, Wikimedia, wikifarm
  - **Zonas:** título, abstract, palabras clave
  - **Servicios:** bases de datos (Scopus), buscadores (Google Scholar), redes sociales (Bibsonomy, CiteULike, CiteSeerX, Zotero) y repositorios de pre-prints (arXiv)
  - **Rango fechas:** de 2001 a 2012

# Estado del arte (II)

**El estado del arte ha quedado dividido en:**

- Autoría y calidad
- Cobertura y sesgos
- Comunidad
- Educación
- GLAM
- Motores wiki
- Predicción y tendencias
- Preservación
- NLP
- Recomendación de tareas
- Semántica
- Vandalismo y spam
- Visualización
- Wikifarms

# EdA: Autoría y calidad

- Los wikis permiten colaborar a **todo el mundo**
  - Hay **controversia sobre calidad** de Wikipedia
- *Nature* (Giles, 2005)
  - De 42 artículos sobre ciencia, **Wikipedia tenía cuatro errores y *Britannica* tres**. Gana *Britannica* pero no tiene tanta ventaja como cabría esperar
- Se han desarrollado **herramientas** para medir calidad de textos
  - **WikiTrust**: colorea las frases que más han durado sin ser modificadas (puede indicar que es correcta)
  - **Article Feedback Tool**: permite informar de errores

# EdA: Cobertura y sesgos

- El desarrollo de Wikipedia (280 idiomas) es **desigual**
  - Algunos con millones de páginas, otros con miles
- (Warncke-Wang et al., 2012)
  - Wikipedia en inglés es de donde traducen las demás
- Se está intentando **atraer a editores** en países en vías de desarrollo
- Poca participación de mujeres en Wikipedia
  - Entre el 5-20% de las aportaciones solamente



# EdA: Comunidad

- Más de **17 millones de usuarios** en Wikipedia
  - **130.000 activos** en el último mes
- Para organizar esta comunidad existen:
  - **Normas y políticas** sobre:
    - Comportamiento, manual de estilo, resolución de conflictos, etc
- (Fordand Geiger, 2012) y (Zhang et al., 2012)
  - La **incorporación de usuarios nuevos es costosa**
    - Curva de aprendizaje alta: interfaz, políticas...

# EdA: Educación

- Se han empleado **wikis en asignaturas**
  - (Rodríguez-Posada et al., 2011; Palomo-Duarte et al., 2012)
- Los alumnos **escriben** los trabajos **colaborativamente** y son evaluados
- El trabajo queda en Internet para su consulta
- Tiene **mayor impacto**
- También, si se cumplen las normas de Wikipedia, los textos pueden acabar siendo incorporados a la enciclopedia

# EdA: GLAM

- Bibliotecas, archivos y museos **dedican mucho presupuesto a digitalizar** materiales
  - Pero logran **poca visibilidad** en buscadores
- (Lally and Dunford, 2007; Danielle Elder and Reilly, 2012)
  - Hay **experiencias para incorporar referencias** en Wikipedia hacia estas entidades culturales, con buenos resultados (**aumentan su tráfico**)
- **Europeana** ha creado un **widget** para exportar referencias hacia Wikipedia

# EdA: Motores wiki

- El motor wiki más conocido es **MediaWiki**
  - Pero **existen más de 100** ([www.wikimatrix.org](http://www.wikimatrix.org))
- Hay poca interoperabilidad entre ellos
  - Cada uno tiene su propia sintaxis
    - Negritas, cursivas...
    - Inserción de imágenes
- El proyecto **WikiCreole** intenta solucionarlo
  - Promueve una sintaxis estándar
  - Están trabajando en ello

# EdA: Predicción y tendencias

- **Wikipedia es actualizada constantemente**
  - Desastres, eventos deportivos, fallecimientos famosos, son incorporados muy rápido
- (Osborne et al., 2012)
  - Se están desarrollando **sistemas que detectan las subidas de actividad** en ciertos temas
- (Márton Mestyán and Kertész, 2012)
  - También con carácter predictivo, por ejemplo: **precedir el éxito en taquilla de películas** en base a la actividad del artículo de Wikipedia

# EdA: Preservación

- Existe miles de wikis en Internet
- Muy pocos ofrecen *backups* públicos
  - Excepciones como Wikipedia, que sí los ofrecen
- Proyectos como **WikiTeam** intentan solucionarlo
  - Ya han extraído datos de +5000 wikis
  - Es posible analizar todo estos datos y extraer conclusiones

# EdA: NLP

- Wikipedia cuenta con **+20 millones de artículos en +280 idiomas**
- Esto es un **corpus lingüístico** enorme
- Se están empleando técnicas para analizarlo
- (Ferron and Massa, 2012)
  - Artículos sobre desastres no son todo lo neutrales que deberían

# EdA: Recomendación de tareas

- Es un **área poco explorada**
- Los usuarios se beneficiarían de saber:
  - Qué artículos requieren ser escritos
  - Cuáles deben ser mejorados
  - Qué imágenes hacen falta
- Hasta ahora todo esto se hace principalmente de forma manual
  - Lo que supone bastante esfuerzo



# EdA: Semántica

- Incrementa la **expresividad de los datos**
  - Permite relacionar conceptos
  - Inferir conocimiento
- (Krötzsch et al., 2006)
  - Se creó la extensión **Semantic MediaWiki** para explotar los datos
- (Aueret al., 2007)
  - **DBpedia** extrae información de Wikipedia de forma estructurada para ser consumida por máquinas

# EdA: Vandalismo y spam

- Cualquiera puede modificar un wiki
  - Incluso usuarios malintencionados
- Se han desarrollado herramientas contra...
  - **Vandalismos**: bots, herramientas semiautomáticas
    - En Wikipedia en español (Rodríguez-Posada, 2009, 2010)
    - En Wikipedia en inglés (Carter, 2008)
  - **Spam**: filtros, captchas, etc

# EdA: Visualización

- La **gran cantidad de datos** de wikis hacen necesario:
  - Técnicas de **visualización**
  - Y herramientas especiales
- Algunos ejemplos incluyen:
  - **HistoryFlow** (Fernanda B. Viégas and Dave, 2004)
  - **StatMediaWiki** (Rodríguez-Posada et al., 2011)
- Y existen muchas otras, incluso una competición llamada **WikiViz**

# EdA: Wikifarms

- Además de Wikipedia, existen multitud de wikis dedicados a temas concretos
- Se organizan en *wikifarms*
  - Wikia
  - EditThis
  - Tropical Wikis
  - ShoutWiki
- Hay muy **pocos estudios** sobre estos conjuntos de wikis
  - Se podrían hacer **análisis comparativos** con Wikipedia

# Cuestiones abiertas

- Algunos ejemplos de cosas a explorar:
  - **Sistemas de recomendación**
  - Análisis de imágenes
  - Estudiar otros idiomas de Wikipedia
    - Se concentra casi todo en Wikipedia en inglés
  - Estudiar las *wikifarms*
  - **Algoritmos** para detectar ataques y vandalismos
  - Recientemente se ha creado **Wikidata**
    - Es un proyecto de datos estructurados
    - Puede ser un éxito como lo fue Wikipedia en su día

# Conclusiones y trabajo futuro

- **WikiPapers** es un **sistema para recopilar literatura sobre wikis** (ya hay +1700 *papers*, +90 *datasets*, +40 herram., +1000 autores)
- Soluciona los problemas que existían en otros enfoques
  - **Tiene mejoras:** estadísticas, gráficas, tablas, enriquecido con semántica
- WikiPapers es un **estado del arte** en continua actualización:
  - Es actualizado si aparecen nuevos papers
    - A diferencia de los **Systematic Literature Reviews** que lo publicas en una revista y son estáticos
- Ejemplo práctico, las **gráficas** de esta presentación se han extraído de WikiPapers:
  - **Y se actualizan automáticamente cuando hay cambios**
- **El proyecto ha tenido buena acogida** entre los investigadores de este campo:
  - Y algunos ya han participado añadiendo publicaciones

# Gracias por su atención



**¿Preguntas?**

<http://wikipapers.referata.com>