

CONSISTENCIA EN LA INDIZACIÓN DE DOCUMENTOS ENTRE INDIZADORES NOVELES

*Isidoro Gil Leiva**

Universidad Politécnica de Valencia. Facultad de Informática. Licenciatura en Documentación.

Resumen: Se describe una serie de ensayos para conocer los índices de consistencia en la indización de diferentes tipos de documentos (artículo de revista, gráfico y fotografía). Para ello se ha trabajado con alumnos a los que se ha dado treinta horas de formación en indización. A falta de trabajar con una muestra mayor, se concluye que los índices de consistencia van disminuyendo progresivamente según se trate de un artículo, de un gráfico o de una fotografía, y que los índices de consistencia logrados por los indizadores noveles no están muy por debajo de los recogidos en la literatura para este tipo de ensayos.

Palabras clave: Indización de documentos. Consistencia en la indización. Índices de consistencia. Indizadores noveles.

Abstract: The paper describes some essays in order to know the consistency of indexation of different kinds of documents (article, graphics, picture). Indexer were trained during 30 hours in indexation methods. It worked with 3 groups of 3 indexers. In conclusion, the consistency levels are lower in cases of graphics and pictures than in articles. In addition, results of those new indexers are similar than obtained in the scientific literature about this topic.

Keywords: Indexation. Consistency. Consistency Indexes. Indexers.

INTRODUCCIÓN

Uno de los elementos que se pueden estudiar para comprobar la calidad de la indización efectuada es la consistencia. La consistencia consiste en la búsqueda de las semejanzas durante la asignación de palabras clave, materias o descriptores a un documento. Cuando las comparaciones se realizan entre el resultado de un indizador en el análisis de un mismo documento en períodos diferentes, se conoce como consistencia intraindizador; en cambio, la confrontación entre el resultado de varios indizadores en el análisis de un mismo documento se denomina consistencia interindizador. La consistencia entre indizadores viene siendo un tema de estudio desde hace bastantes años como se puede comprobar en Zunde y Dexter [1969], Leonard [1977] o Markey [1984], o más recientemente en los trabajos de Iivonen [1990], Sievert y Andrews [1991], Reich y Biever [1991], Tonta [1991], David y Giroux [1995], Iivonen y Kivimaki [1998], Hudon [1999], Gil Leiva [1999, 2001].

En la indización de un documento intervienen diversas variables. Algunas de estas variables son la formación de la persona que analiza el documento y su experiencia en tareas de indización; el dominio de las herramientas empleadas en la indización, en el

* isgil@har.upv.es

caso de su utilización; el conocimiento del ámbito temático en el que se enmarca el documento; las directrices de indización marcadas por el centro; y por último, la variable más importante, cuando se compara la indización en dos unidades, la muestra analizada debe ser representativa del total de documentos. Por tanto, no sería recomendable sacar conclusiones definitivas sobre la consistencia en la indización entre varias bases de datos o unidades de información hasta que no se controlen todas estas variables.

En el trabajo que exponemos a continuación no se han controlado todas estas variables, por lo que los resultados obtenidos no deben tomarse como definitivos, pero sí nos sirven para poseer más información sobre el proceso de aprendizaje en la indización de documentos y el grado de complejidad en la indización según el tipo de documento.

1. MATERIAL Y MÉTODO

En el estudio de consistencia en la indización han participado alumnos de la asignatura Técnicas de indización de la Licenciatura en Documentación de la Universidad Politécnica de Valencia. Una vez finalizada la asignatura y tras recibir aproximadamente treinta horas de formación tanto teórica como práctica sobre indización, los alumnos hicieron un ejercicio práctico consistente en la indización de una serie de documentos utilizando el Tesouro Eurovoc. Los documentos fueron un artículo de revista, un gráfico y una fotografía (ver anexo 1). Los alumnos debían anotar para cada documento indizado en una columna las palabras clave y en otra columna los descriptores obtenidos tras la consulta a Eurovoc.

Se realizaron, por tanto, tres ensayos prácticos para calcular los índices de consistencia entre indizadores: ensayo 1 (artículo de revista); ensayo 2 (gráfico); ensayo 3 (fotografía). En cada ensayo participaron nueve alumnos agrupados en tres grupos, cada grupo compuesto por tres alumnos elegidos al azar. Posteriormente, se hallaron los índices de consistencia de las palabras clave y de los descriptores, comparando los propuestos por un alumno con los otros dos del grupo. Para ello, se empleó la siguiente fórmula:

$$C_i = \frac{T_{co}}{(A+B) - T_{co}}$$

donde,

C_i = Consistencia entre dos indizadores.

T_{co} = Número de términos comunes asignados por los dos indizadores.

A = Número de términos asignados por el indizador A.

B = Número de términos propuestos por el B.

T_{co} = Número de términos comunes asignados por los dos indizadores.

Esta fórmula ha sido empleada para hallar índices de consistencia en la indización de documentos tanto de manera manual como automática en Hooper [1965], Salton y McGill [1983], Lustig y Knorz [1986], Lancaster [1991], Tonta [1991], Silvester, Ge-

nuardi y Klingbiel [1994], Gil Leiva [1999, 2001]. En el anexo 2 se muestra un ejemplo del procedimiento seguido.

2. RESULTADOS Y DISCUSIÓN

A continuación se muestran los datos logrados en los tres ensayos:

Ensayo 1: Consistencia en la indización de artículos de revista

	<i>Consistencia Palabras Clave %</i>	<i>Consistencia Descriptorios %</i>
Grupo 1		
Indizador 1 versus indizador 2	20	· 38
Indizador 1 versus indizador 3	· 27	23
Indizador 2 versus indizador 3	23	23
<i>Media grupo 1</i>	23,3	28
Grupo 2		
Indizador 1 versus indizador 2	· 33	21
Indizador 1 versus indizador 3	· 43	26
Indizador 2 versus indizador 3	· 31	25
<i>Media grupo 2</i>	35,6	24
Grupo 3		
Indizador 1 versus indizador 2	38	· 39
Indizador 1 versus indizador 3	· 54	23
Indizador 2 versus indizador 3	· 46	33
<i>Media grupo 3</i>	46	31,6
<i>Media ensayo 1</i>	34,9	27,8

Ensayo 2: Consistencia en la indización de gráficos

	<i>Consistencia Palabras Clave %</i>	<i>Consistencia Descriptorios %</i>
Grupo 1		
Indizador 1 versus indizador 2	· 57	41
Indizador 1 versus indizador 3	12	· 13
Indizador 2 versus indizador 3	13	· 20
<i>Media grupo 1</i>	27,3	24,6
Grupo 2		
Indizador 1 versus indizador 2	16	· 30
Indizador 1 versus indizador 3	13	· 16
Indizador 2 versus indizador 3	7	· 17
<i>Media grupo 2</i>	12	21
Grupo 3		
Indizador 1 versus indizador 2	· 25	12
Indizador 1 versus indizador 3	12	· 25
Indizador 2 versus indizador 3	· 39	28
<i>Media grupo 3</i>	25,3	21,6
<i>Media ensayo 2</i>	21,5	22,4

Ensayo 3: Consistencia en la indización de fotografías

	<i>Consistencia Palabras Clave %</i>	<i>Consistencia Descriptorios %</i>
Grupo 1		
Indizador 1 versus indizador 2	25	· 40
Indizador 1 versus indizador 3	25	25
Indizador 2 versus indizador 3	17	· 20
<i>Media grupo 1</i>	22,3	28,3
Grupo 2		
Indizador 1 versus indizador 2	22	· 37
Indizador 1 versus indizador 3	· 29	25
Indizador 2 versus indizador 3	28	· 37
<i>Media grupo 2</i>	26,3	33
Grupo 3		
Indizador 1 versus indizador 2	15	· 22
Indizador 1 versus indizador 3	11	· 15
Indizador 2 versus indizador 3	8	· 16
<i>Media grupo 3</i>	11,3	17,6
<i>Media ensayo 3</i>	19,9	26,3

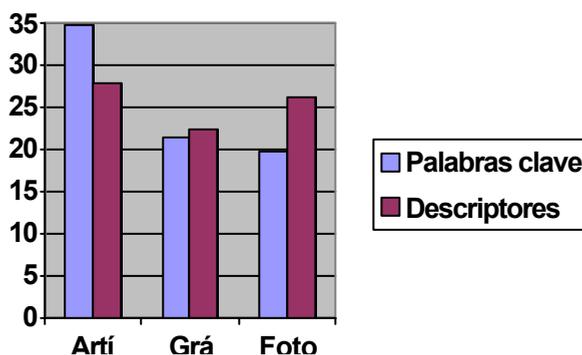
En el análisis de los datos se comprueba que el índice de consistencia en las palabras claves va descendiendo desde el 34,9 % (artículo), 21,5 % (gráfico) al 11,3 % (fotografía). Lo mismo ocurre con la consistencia en los descriptores: 27,8 % (artículo), 22,4 % (gráfico) y 17,6 % (fotografía). El descenso progresivo en la consistencia en las palabras clave es de 10 puntos aproximadamente, mientras que en los descriptores es alrededor de 5. Por otro lado, hemos comprobado cómo dos alumnos que parten de una misma palabra clave llegan a descriptores diferentes y viceversa, es decir, de diferentes palabras claves a un mismo descriptor. Este hecho se debe a varios aspectos, de los que mencionamos solamente tres: a la inexperiencia de los alumnos en las tareas de indización; a que no poseen un conocimiento profundo de la herramienta de indización; y a la propia subjetividad inherente a la indización.

En Gil Leiva [2001] se hallaron índices de consistencia en la asignación de materias en Bibliotecas Públicas del Estado. En la siguiente tabla se muestran los resultados obtenidos en aquel trabajo con los que presentamos aquí:

Ensayo	Índices de consistencia
Materias en Bibliotecas Públicas	Comparación relajada: 46,6 %
	Comparación rígida: 37,7 %
Indizadores noveles	Palabras clave: 34,9 %
	Descriptores: 27,8 %

‘Relajada’ significa que en la comparación entre un encabezamiento y subencabezamiento y otro encabezamiento y subencabezamiento puede coincidir bien el encabezamiento o bien el subencabezamiento; mientras que en la comparación ‘rígida’ a la hora del análisis debe coincidir todo, es decir, el encabezamiento y el subencabezamiento, en el caso de tenerlo. En cambio, los índices de los indizadores noveles corresponden a la media obtenida en el ensayo dedicado a la indización del artículo de revista, puesto que es lo más parecido al material utilizado en el ensayo en Bibliotecas Públicas que fueron libros.

Índices de consistencia por ensayo



CONCLUSIONES

A falta de trabajar con una muestra mayor que la manejada en este estudio, se concluye que:

1. La idea de que el análisis de una imagen es más complejo que un texto parece coincidir con los resultados alcanzados.
2. Los mayores índices de consistencia tanto en palabras clave como en descriptores se han alcanzado en el ensayo del artículo, del gráfico y de la fotografía respectivamente.
3. Los índices de consistencia obtenidos no están muy por debajo de los recogidos en la literatura existente sobre este tipo de ensayos.

BIBLIOGRAFÍA

- DAVID, C. y GIROUX, L. Indexing as problem solving: a cognitive approach to consistency. En Proceedings of CAIS/ACSI 95, 23rd Annual Conference of the Canadian Association for Information Science, 1995, p. 79-89.
- GIL LEIVA, I. La automatización de la indización de documentos. Gijón: Trea, 1999.
- GIL LEIVA, I. Consistencia en la asignación de materias de Bibliotecas Públicas del Estado. Boletín de la Asociación Andaluza de Bibliotecarios, 2001, n° 63, p. 69-86.
- HOOPER, R.S. Indexer consistency test-origin, measurements, results and utilization. Bethesda: IBM Corporation, 1965.
- HUDON, M. An assessment of the usefulness of standardized definitions in a thesaurus through interindexer terminological consistency measurements. University of Toronto, 1999.
- IIVONEN, M. Interindexer consistency and the indexing environment. International Forum on Information and Documentation, 1990, vol. 15, n 2, p. 16-21.
- IIVONEN, M. y KIVIMAKI, K. Common entities and missing properties: similarities and differences in the indexing of concepts. Knowledge Organization, 1998, vol. 25, n 3, p. 90-102.
- LANCASTER, F.W. Indexing and abstracting in theory and practice. London: The Library Association, 1991.
- LEONARD, L.E. (ed.). Inter-indexer consistency studies, 1954-1975: A review of the literature and summary of the study results. University of Illinois, 1977.
- LUSTIG, G., KNORZ, G. AIR/PHYS pilot application project: pilot application of automatic indexing and improved retrieval methods using the PHYS data base (1-30). Karlsruhe: Frachinformationszentrum, Energie Physik Mathematik GmbH, 1986.
- MARKEY, K. Inter-indexer consistency test: a literature review and report of a test of consistency in indexing visual materials. Library and Information Science Research, 1984, vol. 6, 155-177.
- REICH, P. y BIEVER, E.J. Indexing consistency: the input/output function of thesaurus. College and Research Libraries, 1991, vol. 52, n 4, p. 336-342.

- SALTON, G., MCGILL, M.J. Introduction to modern information retrieval. New York: McGraw-Hill, 1983.
- SIEVERT, M.E. y ANDREWS, M.J. Indexing consistency in Information Science Abstracts. Journal of the American Society for Information Science, 1991, vol. 42, n 1, p. 1-6.
- SILVESTER, J.P., GENUARDI, M.T., KLINGBIEL, P.H. Machine-aided indexing at NASA. Information Processing & Management, 1994, vol. 30, n° 5, p. 631-645.
- TONTA, Y. A Study of indexing consistency between Library of Congress and British Library Catalogers. Library Resources and Technical Services, 1991, vol. 35, n° 2, p. 177-85.
- ZUNDE, P. y DEXTER, M.E. Indexing consistency and quality. American Documentation, 1969 july, p. 259-267.

ANEXO 1

INDICE DE DESIGUALDAD POR COMUNIDADES AUTONOMAS

Begoña GARCÍA GRECIANO

1. INTRODUCCION

La relevancia del proceso de igualdad entre comunidades autónomas (CC.AA.) ha tomado especial importancia desde la adhesión de España a la Comunidad Europea, y más de cara a la integración en un mercado único. En esta nota se analiza la tendencia seguida por la evolución de dicho proceso de igualdad entre comunidades autónomas en los últimos treinta años. Para ello, se ha calculado un índice de desigualdad sobre dos magnitudes macroeconómicas: el producto interior bruto por habitante (PIBpc) y la renta familiar bruta disponible por habitante (RFBdp). El índice de desigualdad (*ID*) por comunidades autónomas intenta medir el grado de desviación de las comunidades con respecto a la media nacional, y en qué medida se han reducido las disparidades regionales. Según se realice el cálculo sobre cada uno de los dos indicadores que acabamos de citar, el enfoque del proceso de igualdad entre comunidades tendrá matices diferentes.

En general, el índice de desigualdad se ha calculado del siguiente modo:

$$ID = \sqrt{\frac{\sum_{i=1}^n \left(\frac{Y_i}{\bar{Y}} - 1 \right)^2}{n}}$$

donde: *Y* es el PIB per cápita, o RFBdp per cápita, para cada una de las comunidades autónomas e *Y* es el PIB per cápita, o RFBdp per cápita, para el total nacional, siendo *n* = 17. Todas las variables están medidas a precios corrientes de cada año. El *ID* será igual a cero para máxima igualdad de las comunidades autónomas.

En el gráfico adjunto, se han representado ambos índices calculados según el PIBpc o la RFBdp. El periodo temporal para el producto interior bruto por habitante va de 1960 hasta 1992, y para la renta familiar bruta disponible por habitante, de 1967 a 1992 (no se dispone de datos de renta disponible por CC.AA. con anterioridad a 1967). Los datos de los últimos años se han obtenido de las estimaciones del crecimiento del PIB por comunidades autónomas que viene realizando la Fundación FIES, de las Cajas de Ahorros Confederadas, enlazando las series hacia atrás con los datos de la *Renta Nacional de España*, del Banco Bilbao Vizcaya.

A partir del gráfico, se han realizado dos tipos de análisis. El primero se centra en el estudio de las tendencias seguidas por los índices de desigualdad a lo largo del tiempo, y el segundo es un análisis comparativo de los *ID* calculados en términos de PIBpc versus RFBdp.

2. ANALISIS DE LA TENDENCIA DEL INDICE DE DESIGUALDAD POR COMUNIDADES AUTONOMAS

El gráfico pone de manifiesto un cambio de tendencia en la evolución del índice de desigualdad a partir de 1979, tanto en términos de PIBpc como de RFBdp, lo que nos permite dividir el periodo en dos etapas. Una primera etapa queda marcada por una continuada reducción de las disparidades regionales que se prolonga hasta el año 1979. La segunda etapa, de 1979 a 1992, se caracteriza por un freno en el proceso de igualdad.

Para evaluar las causas que generan este cambio de tendencia entre

las dos etapas, nos debemos remitir a las variaciones de las magnitudes sobre las que se calcula el índice de desigualdad. Dado que los indicadores se miden en términos per cápita, la desigualdad tendrá en cuenta tanto la localización del producto, o la de la renta disponible, como la de la población. Por tanto, desglosemos las causas diferenciando las variaciones del PIB, o RFBdp, de las variaciones de la población.

En primer lugar, las variaciones de la población jugaron un papel importante debido a los movimientos migratorios regionales. Durante los años sesenta y setenta, la población emigró de las comunidades más pobres a las más ricas, aumentando el número de población de aquellas regiones con mayores niveles de producción (renta), lo que haría caer la medida del PIB o RFBdp, en términos per cápita, en las regiones más ricas. Del mismo modo, se registró una reducción de la población en las comunidades más pobres, aumentando así el valor de su producción en términos per cápita. Este factor genera parte de la reducción del índice durante estas dos décadas, reflejando una mayor igualdad por comunidades autónomas. En los años ochenta, el proceso migratorio no sólo fue menor, sino que incluso se registró un movimiento inverso, aunque en menor medida, de retorno a las regiones de origen. Así, en la segunda etapa, de nuevo este efecto puede reflejar parte del freno de la tendencia a la igualdad de las dos décadas anteriores.

Como se ha indicado, otras causas que influyen en el cambio de la tendencia del índice de desigualdad hay que buscarlas en las variaciones del *output* o de la renta familiar disponible, por comunidades. En términos de producción, los cambios estructurales que se producen dentro de cada comunidad autónoma y las variaciones de sus productividades sectoriales afectan directamente a las variaciones del producto final. Así, cabe destacar, por ejemplo, que el proceso de *terciarización* ha podido favorecer la igualdad entre comunidades autónomas. Es decir, dada la



ganancia de peso del sector servicios sobre el valor total del PIB, a escala nacional y por CC.AA. unido a que es el sector más homogéneo por CC.AA., puede haber contribuido a la reducción de las disparidades regionales.

3. EL INDICE DE DESIGUALDAD DESDE EL ENFOQUE DEL PIB POR HABITANTE VERSUS RFBD POR HABITANTE

A partir del análisis comparativo de los índices de desigualdad calculados según el PIBpc versus RFBDpc, el gráfico muestra, primero, que en términos de renta familiar disponible el índice siempre es menor, y segundo, que a lo largo del tiempo las diferencias entre ambos índices se han incrementado.

Como era de esperar, el índice de desigualdad calculado con la renta

familiar bruta disponible per cápita es menor que el calculado con el producto interior bruto per cápita; por tanto, las comunidades autónomas son más iguales en términos de RFBD por habitante.

El propio concepto de estas dos variables macroeconómicas genera estos resultados. El producto interior bruto es el valor de los bienes finales producidos en el interior de un país (región), mientras la renta familiar bruta disponible es la suma de rentas brutas más transferencias menos impuestos directos y cuotas de la seguridad social. Por tanto, estos dos agregados tienen distintos valor y significado económicos.

Evidentemente, en las comunidades con mayores niveles de producción (renta), los impuestos directos sobre las familias y las cuotas de la seguridad social son mayores que en las comunidades con menores niveles de renta. Si a esto le añadimos

que las transferencias que realizan las administraciones públicas en su papel redistributivo tienden a recaer sobre las regiones más pobres, las disparidades regionales serán menores según el concepto de renta familiar disponible, y el índice de desigualdad tendrá valores más pequeños. Por tanto, la acción redistributiva del sector público juega un papel fundamental en el proceso de igualdad entre comunidades autónomas. En este sentido, hay que resaltar los efectos de las políticas regionales no sólo en cuanto a prestaciones sociales (subvenciones, subsidios, etc.), sino también en cuanto a exenciones fiscales de apoyo a determinadas áreas.

Finalmente, a lo largo del tiempo, las diferencias entre los índices de desigualdad del PIBpc versus RFBDpc han aumentado, haciendo que, en los últimos años, la renta familiar disponible sea todavía más igualitaria por CC.AA. que el producto interior bruto. En efecto, el gráfico muestra cómo a partir de 1985 estas distancias se hacen más significativas. Un factor que puede haber contribuido a que se incrementen estas diferencias es que, con anterioridad a 1985, las políticas regionales las centralizaba la Administración central, y posteriormente cuentan también con las actuaciones de los gobiernos autonómicos y de la Comisión de las Comunidades Europeas, que prestan ayudas adicionales, entre las que cabe destacar el Fondo de Compensación Interterritorial y los fondos estructurales de la Comisión de las Comunidades Europeas. Estas ayudas han favorecido la igualdad regional en términos de renta familiar disponible.

4. CONCLUSIONES

Primera. El gráfico ilustrado nos permite hablar de proceso hacia la igualdad entre CC.AA. a lo largo de los últimos treinta años; sin embargo, la intensidad de dicho proceso no es la misma en todo el período. En efecto, en una primera etapa, hasta 1979, se reducen claramente las dis-

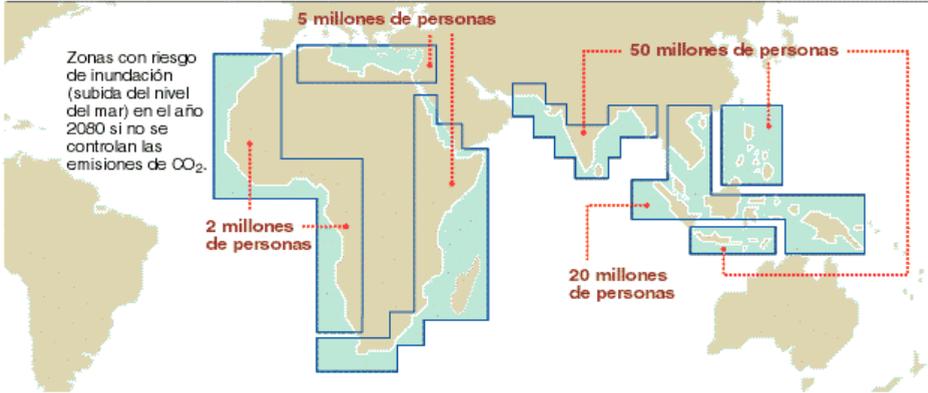
paridades regionales, mientras de 1979 a 1992 se frena el proceso hacia la igualdad tanto en términos del PIB per cápita como de renta familiar bruta disponible per cápita. A partir de 1987, el índice de desigualdad de la RFBdpc es estable. Este cambio de tendencia puede venir en parte explicado por los movimientos migratorios regionales y por los cambios estructurales producidos en cada comunidad autónoma, y que diferencian las dos etapas.

Segunda. El *ID* calculado según la RFBdpc es siempre menor que el calculado según el PIBpc. Es decir, en términos de renta familiar disponible las CC.AA. son más iguales que en términos de producción. El propio concepto de RFBdpc produce este resultado, jugando un papel importante, entre otros, la acción redistributiva del sector público.

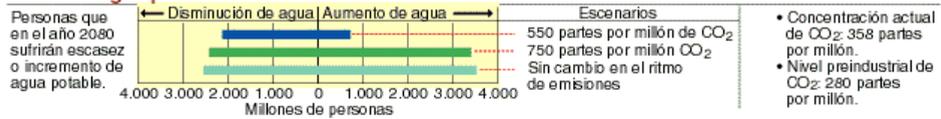
Tercera. Las diferencias entre ambos *ID* se han incrementado a lo largo del tiempo, pero más significativamente a partir de 1985, pudiendo haber contribuido el aumento de las prestaciones sociales que reciben las CC.AA., ya que, a las ayudas de la Administración central hay que sumar las de los gobiernos autonómicos y las de la Comisión de las Comunidades Europeas.

Impactos del cambio climático

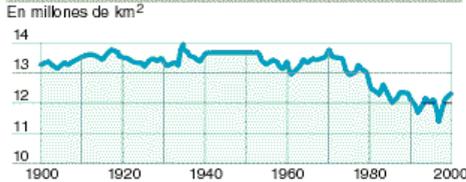
Población afectada por el aumento del nivel del mar



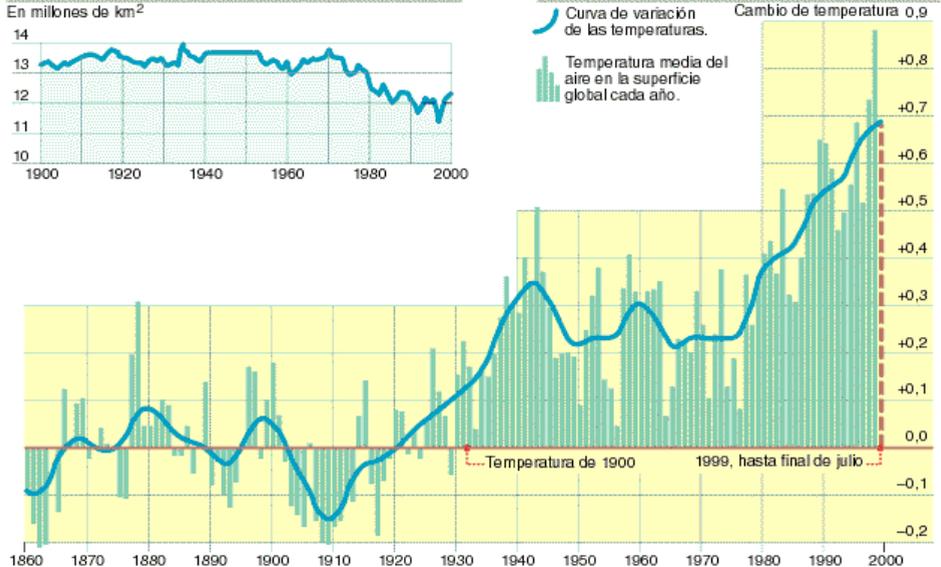
Crisis de agua potable



Evolución de la superficie de hielo en el Ártico

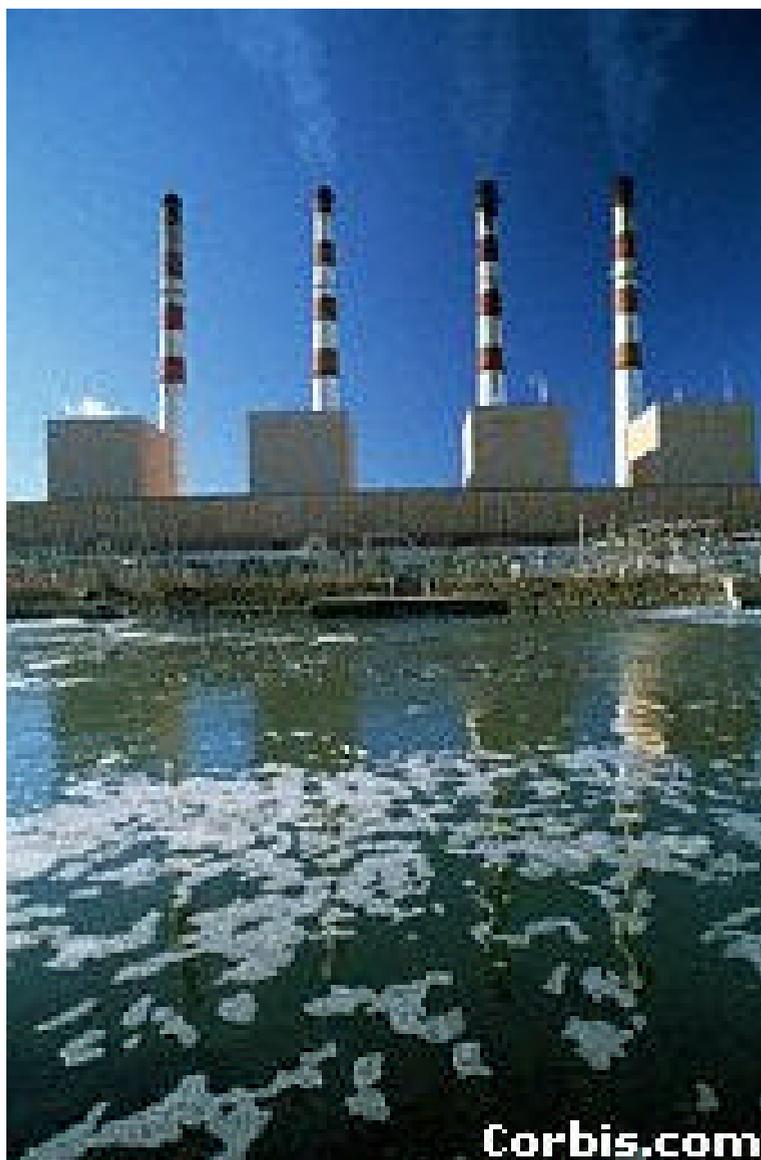


Temperaturas desde la industrialización



Fuente: Hadley Center

A. A. / EL PAÍS



ANEXO 2

*Ensayo 3: Consistencia en la indización de fotografías***Indizador 2:**

PALABRAS CLAVE	DESCRIPTORES
1. Actividad industrial	1. Empresa industrial
2. Chimeneas +	
3. Contaminación +	2. Contaminación industrial +
4. Fábricas	3. Edificio industrial +

Indizador 3:

PALABRAS CLAVE	DESCRIPTORES
1. Chimeneas +	1. Contaminación del agua
2. Contaminación del agua +	2. Contaminación industrial +
3. Industria	3. Emplazamiento industrial +
4. Industrialización	4. Industrialización
5. Países industrializados	5. País industrializado
6. Vertidos industriales	6. Residuo industrial

Si se comparan las palabras claves propuestas por el indizador 2 y el indizador 3 se comprueba que: 1. Los dos apuntan la palabra 'chimeneas', por tanto, 1 término en común; 2. Mientras el indizador 2 señala 'contaminación' el indizador 3 propone 'contaminación del agua', lo que tomamos como una coincidencia del cincuenta por ciento, es decir, 0,5 términos en común. En definitiva, 1,5 términos en común. Del mismo modo, se procede en la comparación de los descriptores.

A continuación, se aplica la siguiente fórmula al indizador 2 (es decir A) y al indizador 3 (esto es B):

$$C_i = \frac{T_{co}}{(A+B) - T_{co}}$$

Palabras Clave $C_i = \frac{1,5}{(4+6) - 1,5} = 17 \%$

Descriptores $C_i = \frac{1,5}{(3+6) - 1,5} = 20 \%$