# Propuesta de un proceso de revisión sistemática de experimentos en Ingeniería del Software

Anna Grimán Padua[1]

Universidad Simón Bolívar, Departamento de Procesos y Sistemas, Apartado Postal 89000, Caracas – Venezuela
agriman@usb.ve

**Abstract.** Las revisiones sistemáticas (RS) han ganado reconocimiento dentro de la Ingeniería de Software (IS). Sin embargo, se cuenta con pocos referentes acerca de cómo conducir una RS en esta disciplina. Los lineamientos existentes muestran una fuerte influencia de Medicina, cuya práctica es mucho más madura. Por ello, tales propuestas no son completamente aplicables a la IS; planteando así necesidades de mejora y adaptación que sólo pueden ser resueltos con un proceso de RS desarrollado específicamente para esta disciplina. El objetivo general de esta investigación es desarrollar un proceso de RS de experimentos adaptado a la IS que cubra las tareas de: *Búsqueda*, que será de especial importancia para la localización de los estudios primarios y también para la definición del objetivo de la revisión; *Extracción de datos*, que tendrá como meta aportar a los datos experimentales de uniformidad y formalidad; y, finalmente, *Síntesis de resultados*, que se enfocará principalmente en la obtención de nuevas evidencias a través del uso de meta-análisis.

**Keywords:** Ingeniería de Software Empírica, experimentos, revisiones sistemáticas, meta-análisis.

## 1 Descripción del problema y estado del arte

Las RS tuvieron su origen en Medicina [20], en donde se definió el concepto como "un estudio de las evidencias que dan respuesta a una pregunta que ha sido claramente formulada por el revisor y que utiliza métodos sistemáticos y explícitos para identificar, seleccionar, y evaluar tales evidencias; así como extraer y analizar los datos proporcionados por éstas" [13]. En la RS pueden o no utilizarse métodos estadísticos, como el meta-análisis. Además de en Medicina, las RS se han aplicado durante años en Educación, Psicología y otros campos de las Ciencias Sociales y de la

[1] Esta investigación se desarrolla bajo la supervisión de Natalia Juristo (natalia@fi.upm.es) y Óscar Dieste (odieste@fi.upm.es) en la Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte 28660, Madrid, España

salud, para la generación de evidencias que soporten la toma de decisiones en su práctica profesional. Como ejemplo se puede mencionar el catálogo de la Cochrane que cuenta con más de 7000 revisiones del área de Medicina. En tales disciplinas existen metodologías reconocidas, tales como [6,7,10,13,18,19], que orientan en la realización de estos procesos y el reporte de sus resultados.

En cambio, las RS se han introducido muy recientemente en la IS. Hasta la actualidad, sólo se han realizado alrededor de una docena de RS en IS; ejemplos de ellas son [3,5,11,12,14,17]. Existen pocas orientaciones sobre cómo realizarlas [1,15] siendo la más popular de ellas la propuesta de Kitchenham [15]. El procedimiento propuesto en [15] está dividido en tres fases: Planificación, Conducción y Reporte de la revisión. En esta propuesta es notoria la influencia de las metodologías de RS de Medicina, tales como [10,13,18], sin embargo dichas metodologías no son necesariamente extrapolables a otras áreas del conocimiento y en particular a la IS. La experiencia recogida en [3,9,] nos ha permitido identificar las características de la IS que la diferencian de Medicina en cómo se conduce una RS:

- La motivación se deriva del interés por conocer el estado de madurez del conocimiento sobre una tecnología (por ejemplo, ¿qué sabemos sobre técnicas de prueba?). Como consecuencia no se puede definir un objetivo con suficiente precisión desde el comienzo de la RS [10], y es muy difícil realizar un desarrollo temprano del protocolo de revisión.
- En relación a la *búsqueda*, vemos que en IS la exhaustividad en las búsquedas es necesaria debido a que se cuenta con un número reducido de experimentos. Debido a que la terminología utilizada en el área no es estándar, el proponer cadenas de búsqueda predefinidas es imposible. La *Síntesis de datos* se ve afectada por la falta de estandarización. Experimentos que analizan los mismos factores y variables respuesta los pueden reportar de manera diferente, creando confusión al agregarlos.
- La *extracción de datos* no es sencilla ya que también se ve afectada por la falta de estandarización del vocabulario y la baja calidad de los reportes.
- Los aspectos de un estudio (por ejemplo, ocultamiento, enmascaramiento, etc.) que influyen en su *calidad* no son claros. Esto que ha resultado evidente en nuestros resultados preliminares, también ha sido ampliamente discutido en el caso de otras disciplinas [2,8,16], en las que tampoco se tiene claro qué aspectos pueden influir directamente en los resultados del estudio.
- El contar con un número muy pequeño de experimentos, que incluyen pocos sujetos, afecta al nivel de certeza de los resultados obtenidos a través de *meta-análisis* tradicional (especialmente si el tamaño del efecto estudiado es muy pequeño).. Además de esto, existen diferentes variables moderadoras que pueden influir en los resultados de cada experimento llevando muchas veces a obtener resultados muy heterogéneos**,** incluso contradictorios.

En las características anteriores podemos observar que más que aplicar metodologías importadas de otras disciplinas se requiere un proceso particular de RS para IS. La ausencia de tal proceso provoca que la Síntesis de resultados experimentales sea muy difícil y que no sea posible generar piezas de conocimiento que soporten adecuadamente la toma de decisiones en esta disciplina.

## 2 Objetivos de la investigación

El objetivo general de esta investigación es *desarrollar un proceso de revisión sistemática adaptado a la IS, que aborde los aspectos inherentes a la búsqueda, extracción de datos, valoración de la calidad, y síntesis de datos*. Para cada una de estas etapas, se han de solventar las limitaciones descritas anteriormente. De este objetivo general se desprenden como objetivos específicos:

- Proponer un mecanismo para establecer la pregunta de revisión en aquellos casos en los que no se cuente con un objetivo específico desde el comienzo del proceso.
- Proponer un mecanismo para localizar de manera óptima experimentos para la RS.
- Desarrollar un procedimiento para obtener los datos experimentales adecuados.
- Determinar qué aspectos de los relacionados con un estudio pueden ser utilizados para valorar la calidad de experimentos en IS.
- Proponer un conjunto de criterios para seleccionar el método más adecuado a la calidad y cantidad de experimentos disponibles para la revisión, además de generar evidencias con un cierto nivel de certidumbre a partir de datos homogéneos.

Una vez planteados estos objetivos preliminares, es posible establecer el enfoque a ser usado en la investigación para el logro de los mismos. El plan de trabajo propuesto, incluye definir un marco teórico de referencia, el análisis de antecedentes, el desarrollo de la propuesta y su validación a través de casos de estudio.

## 3 Propuesta preliminar

Para resolver en nuestra propuesta las limitaciones identificadas en [15], que son producto de las características particulares de la IS, nos hemos enfocado en cinco aspectos fundamentales de mejora/adaptación:

- No forzamos el definir a priori un protocolo de revisión, ya que asumimos que el revisor puede no haber establecido aún con precisión el objetivo de la misma. En tal caso, el objetivo será definido cuando se realice un "mapa" de los experimentos con que se cuenta en IS y los temas tratados (ver actividades 1 a 4 de la Búsqueda).
- Hemos desarrollado heurísticas para optimizar la búsqueda y la selección de las bases de datos bibliográficas más adecuadas (ver la actividad 3 de la Búsqueda).
- La fase de Extracción incluye actividades que pueden ejecutarse de manera secuencial o iterativa, así como heurísticas para resolver los problemas originados en la falta de uniformidad de la información en los reportes experimentales.
- Hemos tomado únicamente el diseño del estudio como un criterio de calidad a ser considerado dentro de los criterios de inclusión/exclusión. Sólo incluiremos estudios del tipo "experimento-controlado". Por otra parte, no incluiremos ningún otro aspecto del estudio (por ejemplo, ocultamiento, enmascaramiento, etc.) dentro de tales criterios, ya que ni siquiera existe consenso respecto de estos aspectos en disciplinas más maduras, como Medicina[2].

[2] Identificar qué aspectos influyen en la calidad del estudio es una investigación que acometeremos en un trabajo futuro.

- Establecemos un procedimiento de agregación para seleccionar el método más adecuado a un grupo de estudios. Además proponemos heurísticas para resolver el problema de la uniformidad de los factores y variables respuesta.

Las premisas anteriores, así como su implementación en nuestra propuesta, son el producto del análisis de los antecedentes localizados, así como del aprendizaje obtenido al aplicar el proceso de RS a un caso de estudio (una RS sobre inspección de software). El proceso propuesto en esta investigación se compone de las siguientes fases: Búsqueda, Extracción de datos, y Síntesis de resultados.

### 3.1 Fase de Búsqueda

Su propósito es localizar de manera óptima los estudios primarios que serán utilizados durante todo el proceso. Pueden tomarse dos caminos, dependiendo del escenario en que se encuentre el revisor: a) el revisor no cuenta con un objetivo específico, sino con una idea general del área de interés (se realizarán las actividades 1 a la 7); b) el revisor cuenta desde el comienzo con un objetivo específico (se realizarán sólo las actividades 5 a 7). Las actividades que componen esta fase:

1. Determinación del tema. Consiste en establecer el área sobre la cual se realizará la revisión (p.e., inspecciones de software).
2. Búsqueda preliminar de experimentos. Consiste en localizar los estudios relevantes respecto del tema establecido en la actividad anterior[3].
3. Descarte de artículos irrelevantes. Mediante un análisis de artículos recuperados en la búsqueda se eliminan los que no corresponden con el tema de la RS.
4. Selección de estudios a agregar. Se realiza una clasificación por sub-tema (o mapa) de los estudios relevantes y se selecciona uno o más grupos de artículos, que llamaremos *grupos candidato a agregación* (p.e., experimentos controlados sobre "comparación de técnicas de lectura de código").
5. Declaración del objetivo de revisión. Definir con precisión el *objetivo de revisión*, la *pregunta de revisión* y los *criterios de inclusión/exclusión de estudios*.
6. Búsqueda y lectura de antecedentes. Se realiza una búsqueda de revisiones previas sobre el mismo objetivo y se decide continuar con la RS o bien parar.
7. Búsqueda refinada. Su propósito es localizar el mayor número posible de artículos relevantes de una manera óptima, realizándose tareas de búsqueda, descarte y selección de estudios con base en los criterios de inclusión/exclusión establecidos[4].

### 3.2 Fase de Extracción de datos

Su objetivo es extraer y codificar los resultados experimentales para la posterior Síntesis a partir del conjunto de estudios empíricos seleccionados en la fase anterior. Las actividades que preliminarmente se han vinculado a la extracción son:

---

[3] Para esta actividad heurísticas particulares de búsqueda de experimentos para IS ha sido propuestas en [4]

[4] Algunas heurísticas que pueden ser aplicadas a esta búsqueda son presentadas en [4], relacionadas con los campos de búsqueda, repositorios, términos de la cadena de búsqueda, y el universo de búsqueda.

1. Lectura superficial. Se realiza una lectura de los experimentos seleccionados en la fase anterior para extraer los datos fundamentales del mismo.
2. Describir y agrupar conceptos. Se establece una semántica común para aquellos aspectos (p.e. los factores, niveles, variables respuesta y constructos) que, siendo equivalentes, son llamados de una manera diferente en los experimentos.
3. Agrupamiento de estudios. Se realizará de manera iterativa para todos los estudios seleccionados como *Grupo candidato a agregación* durante la Búsqueda. La entrada son los datos obtenidos en la lectura superficial y se generan *clusters* que contienen estudios muy compatibles en cuanto al objetivo del experimento.
4. Lectura en profundidad. Los *clusters* obtenidos orientarán la lectura en profundidad de los estudios. El esfuerzo de extracción se concentrará en aquellos *clusters* que sean de interés para los revisores, bien sea por su grado de compatibilidad o por la cantidad de estudios que éstos incluyen.
5. Actualizar grupos. Durante la Lectura en profundidad, se pueden modificar los grupos que han sido establecidos tempranamente.
6. Organizar datos. Preparar los datos experimentales extraídos en *resúmenes* para facilitar su agregación durante la fase de Síntesis de resultados.

### 3.3 Fase de Síntesis de resultados

Se combinan los datos extraídos de los experimentos para generar nuevas piezas de conocimiento a través de meta-análisis. Las actividades que conforman esta fase son:

1. *Seleccionar el método* más adecuado a las características del conjunto de estudios que se desea agregar utilizando los criterios: disponibilidad de datos de entrada, confiabilidad del método, número de estudios y su calidad.
2. Agregar los estudios con el *método seleccionado*, aplicando análisis de sub-grupos para garantizar que los datos agregados poseen una homogeneidad aceptable.
3. Consolidar y clasificar las *evidencias* obtenidas.
4. Generar *evidencias finales* y organizarlas en diferentes niveles de abstracción para obtener un resumen que pueda ser útil para los profesionales.

El trabajo realizado se ha focalizado en el estudio y planteamiento de una propuesta para la Búsqueda, Extracción y Valoración de la calidad. En progreso se encuentra la confección de una propuesta para la Síntesis de resultados. Este trabajo revela la enorme complejidad del proceso de RS, ya que, debido las revisiones exhaustivas y los análisis manuales, involucra un gran esfuerzo y una enorme inversión en coste y tiempo. Pensamos que una manera de aumentar la eficiencia de este proceso, al mismo tiempo que se garantiza la validez de sus resultados, es a través de actividades y heurísticas muy específicas que orienten al revisor y que incluyan instrumentos que permitan documentarlo en profundidad. Por ejemplo, nuestra investigación persigue identificar estrategias de búsqueda de experimentos, como en [4], formularios de adquisición de datos, etc.

En este momento, estamos finalizando la fase de Síntesis de datos. Para finalizar el trabajo nos resta únicamente realizar la validación. Ésta consistirá en la realización de dos casos de estudio, uno con el propósito de comprobar la validez del método propuesto, y el otro, el grado de mejora frente al trabajo seminal de [15].

## Referencias


1. Biolchini, J., Mian, P., Natali, A., Travassos, G.: Systematic Review in Software Engineering. Technical report, COPPE/UFRJ (2005).
2. Concato, J., Horowitz, R.: Beyond randomised vs. observational studies. The Lancet, 363 (2004).
3. Davis, A., Dieste, O., Juristo, N., Moreno, A.: Effectiveness of Requirements Elicitation Techniques: Empirical Results derived from a Systematic Review. In: 14th IEEE International Requirements Engineering Conference RE'06. pp. 176--185. IEEE Press, Minneapolis, USA (2006).
4. Dieste, O. Grimán, A., Juristo, N.: Developing Search Strategies for Detecting Relevant Experiments. Empirical Software Engineering, 1--27 (2007).
5. Dybå, T., Arisholm, E., Sjoberg, D., Hannay, J., Shull, F.: Are two heads better than one? On the effectiveness of pair programming. IEEE Software, 24(6), 10--13 (2007).
6. EPPI-Centre: Core Keywording Strategy: Data collection for a register of educational research. Version 0.9.7. Evidence for Policy and Practice Information and Co-ordinating Centre, London (2002).
7. EPPI-Centre: Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research. Version 0.9.5. Evidence for Policy and Practice Information and Co-ordinating Centre, London (2003).
8. Feinstein, A., Horowitz, R.: Problems with the "evidence" of "evidence-based medicine. Ann. J. Med., 103, 529--535 (1977).
9. Fernández, E.: Aggregation Process with multiple evidence levels for experimental studies in Software Engineering. In: 2nd International Doctoral Symposium on Empirical Software Engineering, pp. 75-81. Universidad Politécnica de Madrid, Madrid, España (2007).
10. Higgins, J., Green, S. (eds.): Cochrane Handbook for Systematic Reviews of Interventions 4.2.6. In The Cochrane Library, 4. John Wiley & Sons, Ltd., UK (2006).
11. Jørgensen, M., Shepperd, M.: A systematic review of software development cost estimation studies. IEEE Transactions on Software Engineering. 33(1), 33--53 (2007).
12. Juristo, N., Moreno, A., Vegas, S.: Reviewing 25 Years of Testing Technique Experiment. Empirical Software Engineering, 9, 7--44 (2004).
13. Khan, K., Khalid, S., Riet, G., Glanville, J., Sowden, A., Kleijnen, J. (eds.): Undertaking Systematic Review of Research on Effectiveness. Technical report, NHS Centre for Reviews and Dissemination, (2001).
14. Kitchenham, B., Mendes, E., Travassos, G. : Cross versus within-company cost estimation studies: A systematic review. IEEE Transactions on Software Engineering. 33(5), 316-329 (2007).
15. Kitchenham, B.: Guidelines for Performing Systematic Literature Reviews in Software Engineering, Version 2.3. Technical report, EBSE (2007).
16. Lawlor, D., Smith, G., Bruckdorfer, K., Kundu, D., Ebrahim, S.: Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence?. The Lancet, 363 (2004).
17. Mendes, E.: A Systematic Review of Web Engineering Research. In: ACM/IEEE International Symposium on Empirical Software Engineering, pp. 99--99. IEEE Press, Noosa heads, Australia (2005).
18. NHMRC: How to review the evidence: systematic identification and review of the scientific literature. Australian National Health and Medical Research Council (2000).
19. Pai, M., McCulloch, M., Colford, J.: Systematic Review: A Road Map Version 2.2. Systematic Reviews Group, http://www.medepi.org/meta
20. Pearson, K.: Report on certain enteric fever inoculation statistics. British Medical Journal. 3, 1243--1246 (1904).